



A University of Sussex DPhil thesis

Available online via Sussex Research Online:

<http://sro.sussex.ac.uk/>

This thesis is protected by copyright which belongs to the author.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Please visit Sussex Research Online for more information and further details

Functional characterisation of *pncr003;2L*,
a small Open Reading Frame gene conserved from
Drosophila to Humans

Emile Gerard Magny

Thesis submitted for the Degree of Doctor of Philosophy

School of life Sciences

University of Sussex

Submitted January 2014

UNIVERSITY OF SUSSEX

EMILE GERARD MAGNY

Thesis submitted for the Degree of Doctor of Philosophy

Functional characterisation of *pncr003;2L*, a small Open Reading Frame gene
conserved from *Drosophila* to Humans

SUMMARY

Small open reading frame genes (smORFs) are a new class of genes, which emerged from the revision of the idea that open reading frames have to be longer than 100 codons to be protein coding and functional. Although bio-informatics evidence suggests that thousands of smORF genes could exist in any given genome, proof of their functional relevance can only be obtained through their functional characterization. This work represents such a study for a *Drosophila* smORF (*pncr003;2L*), which was initially misannotated as a non-coding RNA because of its lack of a canonical long open reading frame. Here I show that *pncr003;2L* codes for two small peptides of 28 and 29 aa, expressed in somatic and cardiac muscles. After generating a null condition for this gene, I use the adult *Drosophila* heart as a system to assess the function of *pncr003;2L*. With this system, I show that the small *pncr003;2L* peptides regulate heart contractions by modulating Ca^{2+} cycling in cardiac muscles, with either lack or excess of function of these peptides leading to cardiac arrhythmias, and abnormal calcium dynamics. Finally, through an extensive homology study, I show that these small peptides share a great amount of structural and functional homology with the peptides encoded by the vertebrate smORFs *sarcolipin* (*sln*) and *phospholamban* (*pln*), which act as major regulators of the Sarco-Endoplasmic Reticulum Calcium ATPase (SERCA), the channel responsible for calcium uptake into the ER following muscle contraction.

These results highlight the importance of the *pncr003;2L* smORF and the *Drosophila* system, for the study of cardiac pathologies, but most importantly, they show that this family of peptides, conserved across evolution, represent an ancient system for the regulation of calcium trafficking in muscles. This work corroborates the prevalence, and relevance of this novel class of genes, and shows that closer attention should be given to smORFs in order to determine the full extent of their biological contribution.

DECLARATION

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signed.....

ACKNOWLEDGEMENTS

I would like to thank my D.Phil. supervisor, Prof. Juan Pablo Couso, for giving me the opportunity to be a part of his lab, for his guidance, and most of all, for his understanding and support throughout this process. Juan Pablo has been a source of knowledge and inspiration throughout my D.Phil., and I hope that someday I will be able to reflect this in my scientific career.

I need to thank Jose Ignacio Pueyo-Marques, who has been an incredible driving-force for this project, producing constructs, fly lines, and biochemical experiments, and whose support, in and outside the lab, has been invaluable to me. I would like to thank my co-supervisor, Claudio Alonso, for his good advice, Sarah Bishop for all of her help and advice with the genetics aspects of this project, Frances Pearl for her computational work, and Jeremy Niven for his great work with the intracellular action potential recordings, and for being always available for an interesting discussion. I would especially like to thank Rose Phillips, for doing such a great job as our lab technician, and for keeping everything under control in the lab. I want to thank Julie Aspden, for her friendship, support, and for her accurate insights and comments regarding this work, but most of all, for the encouragement and positivity that her presence brings to the lab. I want to thank my friends and colleagues Unum Amin, Ali Mumtaz, Chris Sampson, Ying Chen, Joao Osorio, Pedro Patraquim, Casandra Villava, Wan Liu, and Richard Kaschula, for their friendship, their encouragement, the good laughs, and for all the lively discussions. My thanks also go to former lab colleagues; Miguel Angel Cespedes, for all his help and advice, particularly with the mutagenesis strategies, and “Uncle” John Chesebro, for being a great friend, and for his amazing cakes. Many thanks to Roger Phillips and Julian Thorpe for their help with the microscopy aspects of this project. And thanks as well as to the CONACYT, and the University of Sussex for funding my research.

Finally, I want to thank Gemma Hopley, for her unconditional love and support throughout this process, for her careful proofreading of this thesis, and for being such a great mum to our two children, Etienne and Sébastien, whom I thank for bringing immense joy, and a sense of balance to my life. This work is dedicated to them

TABLE OF CONTENTS

| | |
|---|----|
| Chapter I - General Introduction..... | 1 |
| 1- Small Open Reading Frame genes (smORFs); what are they and why are they so interesting?..... | 1 |
| 1.1- Genomic annotation methods favour canonical genes over smORFs. | 1 |
| 1.2- <i>tarsal-less</i> : The smallest protein coding gene, plays a big role in <i>Drosophila</i> development. | 7 |
| 1.3 - Targeted computational approaches predict the existence of thousands of smORFs in several organisms. | 9 |
| 1.4 - Assessing the function of new smORFs | 14 |
| 2-Using <i>Drosophila melanogaster</i> as a system to functionally characterise a novel smORF..... | 15 |
| 2.1- <i>Drosophila melanogaster</i> is a good model to study the function of a novel smORF | 15 |
| 2.2-The putative noncoding RNA gene <i>pncr003;2L</i> is a very interesting candidate for this study..... | 17 |
| 2.3-Objectives..... | 20 |
| Chapter II - Materials and methods..... | 23 |
| 2.1-Fly strains. | 23 |
| 2.2-Gamma ray mutagenesis..... | 23 |
| 2.3-RNA-extraction..... | 24 |
| 2.4-DNA-extraction..... | 24 |
| 2.5-cDNA synthesis. | 24 |
| 2.6-Polymerase Chain Reaction (PCR). | 25 |
| 2.7 -List of primers:..... | 26 |
| 2.8- <i>Mhc</i> exon sequencing. | 29 |
| 2.9-Agarose gel electrophoresis..... | 29 |
| 2.10-Minipreparation of plasmid DNA. | 30 |
| 2.11-Plasmid construction. | 30 |
| 2.12-Adult fly, and larvae dissections..... | 31 |
| 2.13-Indirect Flight Muscle sarcomere length measurements..... | 32 |
| 2.14-Simple motility assay. | 32 |

| | |
|--|----|
| 2.15-Flight assay. | 33 |
| 2.16-Heart Video recordings and period measurements. | 33 |
| 2.17-Calcium fluorescent recordings. | 34 |
| 2.18-Intracellular action potential recordings..... | 35 |
| 2.19- <i>In situ</i> hybridisation..... | 36 |
| 2.20-Immunofluorescence. | 37 |
| 2.21-Bioinformatics search for homologues..... | 37 |
| 2.22-Transmission Electron Microscopy..... | 38 |
| | |
| Chapter III - Characterisation of the gene sequence, transcript expression, and translation of <i>pncr003;2L</i> | 39 |
| 1- Introduction: | 39 |
| 2- Results: | 40 |
| 2.1- <i>pncr003;2L</i> codes for two small open reading frames with an optimal translation context..... | 40 |
| 2.2- Expression of the <i>pncr003;2L</i> transcripts. | 44 |
| 2.3- Translation of the <i>pncr003;2L</i> peptides. | 51 |
| 2.4- Subcellular localisation of the <i>pncr003;2L</i> peptides <i>in situ</i> | 55 |
| 3- Discussion:..... | 63 |
| | |
| Chapter IV - The <i>pBac {WH}F02056</i> insertion, and its use to generate null mutants for <i>pncr003;2L</i> | 65 |
| 1- Introduction: | 65 |
| 2- Results: | 67 |
| 2.1- The <i>pBac {WH} F02025</i> line is hypomorphic for <i>pncr003;2L</i> and has a specific muscle phenotype. | 67 |
| 2.2- The phenotype associated with the <i>pBac {WH} F02025</i> line is enhanced by a <i>Myosin heavy chain</i> haploinsufficiency..... | 68 |
| 2.3- Is there a strong genetic interaction between <i>pncr003;2L</i> and <i>Mhc</i> , or are these phenotypes produced by an associated <i>Mhc</i> allele? | 76 |
| 2.4- A <i>pBac</i> transposase-mediated reversion of <i>F02056</i> restores <i>pncr003;2L</i> expression but still presents the short sarcomere phenotype. | 80 |
| 2.5- Mapping of the putative <i>Mhc</i> associated allele. | 83 |
| 2.6- A small intronic deficiency affecting the alternative splicing of <i>Mhc</i> could be responsible for the <i>Mhc^{F02056}</i> allele. | 88 |

| | |
|---|-----|
| 2.7- Using the <i>F02056</i> insertion to generate a small, specific, FRT recombination-mediated deficiency. | 94 |
| 2.8- Using the <i>F02056</i> insertion to generate a directed γ -ray genomic lesion..... | 98 |
| 2.9- Generation of <i>pncr003;2L</i> null backgrounds free of the <i>Mhc^{F02056}</i> allele. | 111 |
| 3- Discussion..... | 118 |

| | | |
|--|--|-----|
| Chapter V- Using the <i>Drosophila</i> adult heart as a system for the phenotypical characterisation of <i>pncr003;2L</i> | | 120 |
| 1- Introduction: | | 120 |
| 2- Results: | | 122 |
| 2.1- <i>pncr003;2L</i> null flies do not display any morphological abnormalities in the adult heart. | | 122 |
| 2.2- Recordings of live beating hearts show that <i>Df pncr003;2L</i> flies display arrhythmic heart contractions. | | 126 |
| 2.3- <i>pncr003;2L</i> rescues the arrhythmicity phenotype | | 130 |
| 2.4- <i>pncr003;2L</i> mutants present abnormalities in their intracellular action potential recordings. | | 139 |
| 2.5- <i>pncr003;2L</i> influences calcium levels during heart contraction. | | 142 |
| 3. Discussion:..... | | 149 |

| | | |
|---|--|-----|
| Chapter VI - Identification of the <i>pncr003;2L</i> smORF as a functional homologue of the vertebrate Sarcolipin / Phospholamban family of regulators of the sarcoendoplasmic reticulum Ca ²⁺ -ATPase (SERCA). | | 152 |
| 1- Introduction: | | 152 |
| 2- Results: | | 153 |
| 2.1-A BLAST search identifies homologue sequences for <i>pncr003;3L</i> in dipterans. | | 153 |
| 2.2- The incorporation of secondary structure comparison, using the PHYRE2 homology search engine identifies the vertebrate <i>sarcolipin</i> smORF as a putative homologue for <i>pncr003;2L</i> | | 157 |
| 2.3- A BLAST search using a phylogenetic consensus sequence between Dipteran <i>pncr003;2L</i> sequences and human Sarcolipin identifies homologues throughout the arthropoda phylum. | | 161 |
| 2.4- The <i>pncr003;2L</i> peptides co-localise, and interact genetically with <i>Ca-P60A</i> , the <i>Drosophila melanogaster</i> homologue of <i>SERCA</i> | | 166 |

| | |
|---|---------|
| 2.5- The vertebrate Sln and Pln peptides partially recapitulate the function of the Scl peptides in flies. | 171 |
| 3- Discussion..... | 176 |
| 3.1 A phylogenetic analysis supports the homology between Scl and Pln/Sln..... | 176 |
| 3.2 Evidence supporting the physical interaction between Scl and Ca-P60A | 177 |
| 3.3 Experimental evidence supporting the functional relation between Scl and Ca-P60A, and the functional homology between Scl Pln/Sln | 178 |
| 3.4 Conclusion..... | 180 |
| Chapter VII - General Discussion..... | 182 |
| 7.1 The Functional homology between Scl Sln and Pln..... | 182 |
| 7.1a Sln and Pln are reversible inhibitors of SERCA, regulated by their phosphorylation state. | 182 |
| 7.1b Regulation of Ca-P60A by a β -adrenergic-like pathway? | 183 |
| 7.1c Mechanistic differences between the Pln and Sln inhibition of SERCA. | 185 |
| 7.1d <i>Drosophila</i> as a model for heart disease | 188 |
| 7.2 Implications of this research on the field of smORFs | 190 |
| 7.2a The potential of smORFs..... | 190 |
| 7.2b putatively non-coding RNAs could represent a rich source of smORFs | 190 |
| 7.2 c Contributions of this work as a case study for the functional characterisation of smORFs..... | 191 |
| References..... | 195 |
| Annexes | 203 |
| Appendix..... | 213 |

LIST OF FIGURES AND TABLES

| | |
|--|----|
| Figure 1.1: Stark contrast between number of short ORFS and annotated genes in the yeast genome | 4 |
| Figure 1.2: Size distribution of annotated genes in humans and <i>Drosophila melanogaster</i> | 6 |
| Figure 1.3 tarsal-less, a smORF gene conserved across arthropods encodes four short peptides of 11-33 amino acids. | 8 |
| Figure 1.4: Size distribution of different pools of predicted smORFs in <i>Drosophila melanogaster</i> | 13 |
| Table 1.1: The conservation scores of a small ORF encoded by pncr003;2L are very similar to those of translated tal AA peptide | 19 |
| Figure 3.1: pncr003;2L, contains two putative smORFs with an optimum context of translation. | 44 |
| Figure 3.2: Transcriptional expression of <i>pncr003;2L</i> | 47 |
| Figure 3.3: The pncr003;2L transcripts are expressed in the Indirect Flight muscles, and show tissue specific expression. | 50 |
| Figure 3.4: Cloning strategy to assess the translation of the <i>pncr003;2L</i> ORF A and ORF B peptides. | 53 |
| Figure 3.5: ORF A and ORF B translation in cultured cells and muscles..... | 54 |
| Figure 3.6: ORF B is translated from the B isoform in situ, and has a similar expression and localisation as ORF A..... | 57 |
| Figure 3.7: The pncr003;2L peptides localise to the dyads. | 60 |
| Figure 3.8: The dyad is at the core of calcium regulation and muscle contraction..... | 62 |
| Figure 3.9: The subcellular localisation of pncr003;2L ORF A and ORF B is not artifactual..... | 62 |
| Figure 4.1: Diagram representing the genomic landscape and genetic deletions surrounding <i>pncr003;2L</i> , and the structure of the <i>pBac{WH}cF02056</i> insertion. | 70 |
| Table 4.1: Initial characterisation of the motility of different genetic conditions affecting pncr003;2L. | 71 |

| | |
|--|-----|
| Figure 4.2: The <i>pBac{WH}F02056</i> insertion gives rise to a hypomorphic condition for <i>pncr003;2L</i> , and to a short sarcomere phenotype. | 74 |
| Figure 4.3: The short sarcomere phenotype associated with the <i>F02056</i> insertion seems to be enhanced by a <i>Mhc</i> haploinsufficiency. | 76 |
| Figure 4.4: The short sarcomere phenotype enhancement is caused by a <i>Mhc</i> haploinsufficiency. | 79 |
| Figure 4.5: The short sarcomere phenotype is independent of the <i>F02056</i> insertion. | 82 |
| Figure 4.6: Diagram representing the homologous recombination protocol implemented to map the putative <i>MhcF02056</i> allele. | 86 |
| Figure 4.7: One out of four <i>chif F02056</i> recombinants has lost the putative <i>MhcF02056</i> allele. | 88 |
| Figure 4.8: Molecular mapping of the <i>MhcF02056</i> allele by sequencing the IFM specific <i>Mhc</i> exons. | 92 |
| Figure 4.9: A specific small intronic sequence deletion in <i>Mhc</i> may be the cause of the <i>MhcF02056</i> allele. | 93 |
| Figure 4.10: Generation of an FRT-mediated specific deficiency removing the <i>pncr003;2L</i> locus. | 98 |
| Figure 4.11: Generation of a Gamma-ray induce a deletion targeting the <i>pncr003;2L</i> locus. | 100 |
| Figure 4.12: genetic and molecular mapping of the Gamma-ray mutants targeting the <i>pncr003;2L</i> locus. | 104 |
| Figure 4.13 As trans-heterozygous, the deficiencies <i>Df γ-ray 6</i> and <i>Df 12</i> , give rise to an homozygous viable null condition for <i>pncr003;2L</i> : | 105 |
| Figure 4.14: <i>pncr003;2L</i> null flies have no structural abnormalities in their muscle organisation. | 110 |
| Figure 4.15 Quantitative flight assays indicate that <i>pncr003;2L</i> null flies can fly normally: | 111 |
| Figure 4.16: Diagram representing the method followed to rescue the <i>CG31739</i> gene, in the <i>Df(2L)12</i> background. | 115 |
| Figure 4.17: Male recombination protocol to remove the <i>MhcF02056</i> allele from the <i>Df γ-ray 6</i> genetic background. | 117 |

| | |
|--|-----|
| Table 4.2: Different genotypes recovered from the Male recombination protocol implemented to remove the <i>MhcF02056</i> allele from the <i>Df pncr003;2L</i> genetic background | 117 |
| Figure 5.1: The <i>Df pncr003;2L</i> null flies show no structural or morphological defects in heart muscles. | 126 |
| Figure 5.2: The hearts of <i>Df pncr003;2L</i> mutants present heart arrhythmias.. | 130 |
| Figure 5.3: The arrhythmias presented by <i>Df pncr003;2L</i> are specific to <i>pncr003;2L</i> and independent from the <i>mhcF02056</i> allele | 135 |
| Figure 5.4: The arrhythmias presented by <i>Df pncr003;2L</i> are corrected with different <i>pncr003;2L</i> expression constructs. | 137 |
| Figure 5.5: <i>pncr003;2L</i> excess of function also leads to heart arrhythmia..... | 139 |
| Figure 5.6: <i>pncr003;2L</i> mutants have abnormal action potential patterns. | 142 |
| Figure 5.7: <i>pncr003;2L</i> null mutants present Ca^{2+} transients with higher amplitudes during heart contraction..... | 146 |
| Figure 5.8: mutant rescues and over-expression effect of <i>pncr003;2L</i> in calcium transients..... | 148 |
| Figure 6.1: Initial analysis of sequence conservation and structure of the <i>pncr003;2L</i> peptides..... | 156 |
| Figure 6.2 The PHYRE2, structural and sequence homology search engine identifies <i>Sln</i> as an <i>pncr002;2L</i> homologue | 160 |
| Figure 6.3: A tBLASTn search using the phylogenetic consensus between the <i>pncr003;2L</i> peptides and <i>Sln</i> , identifies intermediate homologues..... | 164 |
| Figure 6.4: The <i>Sln</i> and <i>pncr003;2L</i> smORFs seem to have a common bilaterian ancestor. | 165 |
| Figure 6.5: The <i>pncr003;2L</i> peptides co-localise with <i>Ca-P60A</i> , the <i>Drosophila</i> SERCA homologue, in the dyads..... | 168 |
| Figure 6.6: <i>pncr003;2L</i> interacts genetically with <i>Ca-P60A</i> | 171 |
| Figure 6.7: The <i>pncr003;2L</i> peptides co-localise with <i>Sln</i> and <i>Pln</i> in the dyads..... | 173 |
| Figure 6.8: Vertebrate <i>Sln</i> and <i>Pln</i> peptides partially recapitulate the function of the <i>Scl</i> peptides. | 176 |
| Figure 7.1: The inhibition of SERCA2a in cardiac muscles by the <i>Sln</i> and <i>Pln</i> peptides is regulated by the β -adrenergic in vertebrates..... | 193 |

| | |
|--|-----|
| Figure 7.2: Possible conserved phosphorylation sites in the Scl family of peptides | 194 |
| Annex 1: List of currently annotated genes coding for peptides under 30 aa long in <i>Drosophila melanogaster</i> | 203 |
| Annex 2: Genetic protocols followed for the generation of the Specific FRT-mediated deficiency and for the pBac{WH}F02056 reversion..... | 205 |
| Annex 3: DNA alignments of <i>pncr003;2L</i> , from the pBac{WH}F02056 lines, and <i>pncr003;2L</i> FS construct. | 208 |
| Annex 4: The Scl peptides can be bioinformatically docked into Ca-P60A, similarly as the Vertebrate Sln and Pln peptides into SERCA..... | 210 |
| Annex 5: Sln and Pln inhibit the activity of SERCA, and their mutants produce cardiac calcium transients comparable to those of Scl null mutants. | 212 |

LIST OF ABBREVIATIONS

| | |
|-------------|---|
| aa | Amino acid |
| AP | Action potential |
| <i>b</i> | <i>black</i> |
| BLAST | Basic local alignment search tool |
| bp | Base pairs |
| <i>cn</i> | <i>cinnabar</i> |
| CAI | Codon adaptation index |
| CaM | Ca ²⁺ calmodulin |
| CaMKII | Ca ²⁺ calmodulin dependent Kinase II |
| cAMP | cyclic adenosine 3'5'-monophosphate |
| Ca-P60A | <i>Drosophila sarcoendoplasmic reticulum ATPase</i> |
| cDNA | Complementary DNA |
| CDS | Coding DNA sequence |
| <i>Chif</i> | <i>chiffon</i> |
| CICR | calium mediated calcium release |
| DAPI | 4',6-diamidino-2-phenylindole |
| <i>Df</i> | Genomic deficiency |
| DGRC | <i>Drosophila</i> genomics resource centre |
| DNA | Deoxyribonucleic acid |
| EST | Expressed sequence tag |
| FH | N-terminal FLAG-Hemagglutinin |
| FLP | Flippase |
| FRT | Flippase recombination site |
| FS | Frame-shift |
| GCaMP3 | GFP-CaM chimeric protein, genetically encoded Ca ²⁺ reporter |
| GFP | Green fluorescent protein |
| IFM | Indirect flight muscle |
| Ka/Ks | Ratio of synonymous/non-synonymous nt changes |
| Kb | Kilo base-pairs |
| lincRNA | long intergenic non-coding RNA |
| <i>Mhc</i> | <i>Myosin heavy chain</i> |
| <i>MYH</i> | <i>Myosin heavy chain</i> |
| Myr | Million years |
| NCX | Na ⁺ /Ca ²⁺ exchanger channel |
| nt | Nucleotides |
| OA | Octopamine |
| ORF | Open reading frame |
| pBac | Piggy-back transposon |
| PCR | Polymerase chain reaction |
| PHYRE2 | Protein homology/ analogy recognition engine |

| | |
|-------------------|--|
| PKA | Protein kinase A |
| Pln | Phospholamban |
| PMCA | Plasma membrane calcium ATPase |
| <i>pncr003;2L</i> | <i>putatively non-coding RNA 003 in chromosome 2L</i> |
| RFP | Red fluorescent protein |
| RNA | Ribonucleic acid |
| RNAi | RNA interference |
| RNA-seq | RNA deep sequencing |
| <i>Rp-49</i> | <i>Ribosomal protein 49</i> |
| RT-PCR | Reverse-transcriptase polymerase chain reaction |
| RV2 | Revertant of pBac{WH}F02056 |
| RyR | Ryanodine receptor |
| Scl | Sarcolamban |
| SER | sarcoendoplasmic reticulum |
| SERCA | Sarcoendoplasmic reticulum Ca ²⁺ ATPase |
| Sln | Sarcolipin |
| smORF | small Open Reading Frame gene |
| <i>tal</i> | <i>tarsal-less</i> |
| tBLASTn | BLAST on translated nucleotide data-base with a protein sequence query |
| TEM | Transmission electron microscopy |
| <i>tin</i> | <i>tinman</i> |
| <i>TpnI</i> | <i>Troponin I</i> |
| UAS | Upstream activator sequence |
| UTR | Untranslated region |
| <i>w</i> | <i>white</i> |
| wt | wild-type |

Chapter I - General Introduction

1- Small Open Reading Frame genes (smORFs); what are they and why are they so interesting?

1.1- Genomic annotation methods favour canonical genes over smORFs.

The basic rules of phenotypic inheritance between a living organism and its progeny were addressed for the first time by Gregor Mendel a century and a half ago. Since then, every branch of biology has pursued, in one way or another, the identification of the molecular mechanisms involved in the transmission of the inheritance of genetic information, and their “phenotypic impact” in organisms. The ultimate aim of this work is to contribute to this very cause, by unveiling the importance of small open reading frames genes (smORFs), one such component, which until now has been neglected.

Most people would agree that the most important component of this inheritance mechanism is the gene. But what is a gene? At its origin, because of the lack of knowledge regarding its biological nature, the term had an abstract nature. It was coined by Wilhelm Johannsen in 1909 as “the special conditions, foundations and determiners which are present in the gametes in unique, separate and thereby independent ways by which many characteristics of the organism are specified” [1,2]. Today, after many major landmark achievements, which have allowed us to understand the molecular nature of entire genomes, because of the dizzying complexity with which we have been confronted, some of which will be discussed in the following sections, it is still tricky to precisely define what a gene is. For a matter of simplicity, and following the established central dogma of molecular biology stating that genomic information flows from DNA to RNA to protein, we could define, for the moment, a canonical gene as the DNA sequence containing the code necessary to generate a protein. This protein coding sequence is known as DNA coding sequence (CDS), when it is known to encode for a protein, or as an open reading frame (ORF), referring to the stretch of nucleotide sequence, which starts with a translational start codon (most commonly an ATG triplet although alternative start codons exist) and finishes with a stop codon.

Such genes have commonly been annotated by large databases / consortia, such as the National Centre of Biological Information (NCBI) or the European Molecular Biology Laboratory project known as Ensembl, with the ultimate aim of identifying and cataloguing all the genes within any given genome. Recently, the Encyclopedia of DNA Elements (ENCODE for humans, or MODENCODE for model organisms) consortia, have gone beyond gene annotation, having taken the task of annotating not only genes, but all coding elements within the human, and model organisms genomes; although this dissertation will touch on such other elements, the focus will remain on genes, as defined earlier. The process of gene annotation usually integrates computational *de novo* gene prediction, which requires the use of sophisticated algorithms often referred to as “gene finders”, and pair-wise alignment to the genome of experimentally supported sequences; usually, these are protein sequences obtained from publications describing the function of particular genes, or from libraries of complementary DNA (cDNA), expressed sequence tags (ESTs), or more recently from RNA sequencing (RNA-seq) reads.

The annotation of *de novo* ORFs with computational prediction methods and, in general, the accurate annotation of genes, is not at all trivial. Allen *et al.* summarise this in their 2007 *Genome research* manuscript [3] pointing out that the difficulties in creating accurate annotations arise for several reasons: “Sometimes the evidence for a gene is weak, consisting of just one gene prediction but no sequence homology, or just a single expressed sequence tag (EST) match. In other cases, the evidence is plentiful but contradictory: Different gene finders and protein sequence alignments may indicate many overlapping candidate genes, and more than one of these models may in fact be correct“. Importantly, they emphasise how time consuming the process of gene annotation can be, especially when the complexity of the prediction/evidence is such that human curation is required.

The already challenging process of gene annotation is even more difficult for ORFs with small sizes, and as a consequence many genes with small ORFs (small ORFs belonging to a gene, and from which a functional peptide is produced, will hereby be referred as smORFs) may have escaped annotation. Basrai *et al.*[4], who argue in favour of this view, plotted a histogram representing the sizes of all ORFs in the yeast genome, superimposed over a histogram representing the sizes of all its annotated genes (Figure 1.1). What this plot shows is that although an immense number of short ORFs

exist in the yeast genome (they report 260,000 ORFs less than 100 codons long), the number of annotated genes encoding small ORFs is relatively very small (~100 genes). Although it is possible that genes with small ORFs may be simply sparse, several factors point to a scenario where they may be plentiful but misannotated. First of all, their short sizes render most computational prediction methods inaccurate. Those methods are usually based on an assessment of the differences in nucleotide composition between coding and noncoding sequences, assessing parameters such as codon usage, sometimes referred to as codon adaptation index (CAI) [5]. CAI calculations take advantage of the fact that in different species codon usages are more or less biased as some amino acids appear to be preferentially encoded by certain codons rather than others.

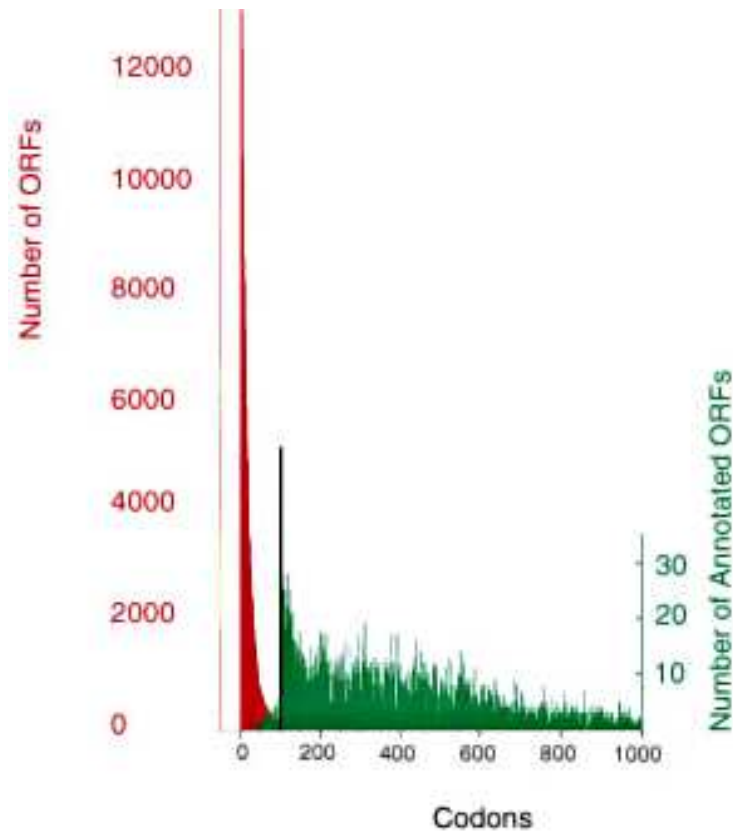


Figure 1.1

Figure 1.1: Stark contrast between number of short ORFs and annotated genes in the yeast genome.

Size distributions of the total number of ORFs encoded in the *S. cerevisiae* genome (red) compared to the total number of annotated genes in the Saccharomyces Genome Database, as published in Basrai et al. (1997) [4]. The black line represents the 100 amino acids size used as a cut-off for many genome annotation projects.

Although these statistical analyses are very effective for distinguishing long coding sequences, they become progressively less useful as ORF length decreases [4]. Furthermore, in order to minimize the number of false positives, these algorithms also integrate a series of other parameters, such as presence of splice sites, promoter regions, translational start/stop sites, and poly-adenylation signals, but this approach also increases the chances for true ORFs not to be detected [6]. Small ORFs seem to be particularly susceptible to false negative predictions, as it is difficult to distinguish the relatively few biologically meaningful sequences, amongst the large number of artefact ORFs present in the genome by pure chance [4,7,8]. The small sizes of their protein coding region, also make smORFs less likely to be targeted by conventional random mutagenesis, which gives to these sequences fewer opportunities to have valuable experimental support. Because of these issues, only genes coding for ORFs of at least 100 contiguous codons were designated for annotation in the yeast genome [9]. Although this decision makes sense since most of the 260,000 yeast ORFs could be artefacts, and it would be much more difficult and time consuming to accurately annotate them, it may have led to the loss of a considerable amount of smORFs coding for peptides with important biological functions. Importantly, this problem is not specific to yeast; the same 100 aa cut-off has been applied to gene annotation in mammals [10], and the size distribution of annotated genes in humans and fruit flies (*Drosophila melanogaster*) show a very similar steep drop in the number of genes coding for proteins under 100 aa long (Figure 1.2), although both organisms have an equally overwhelming amount of small ORFs in their genomes.

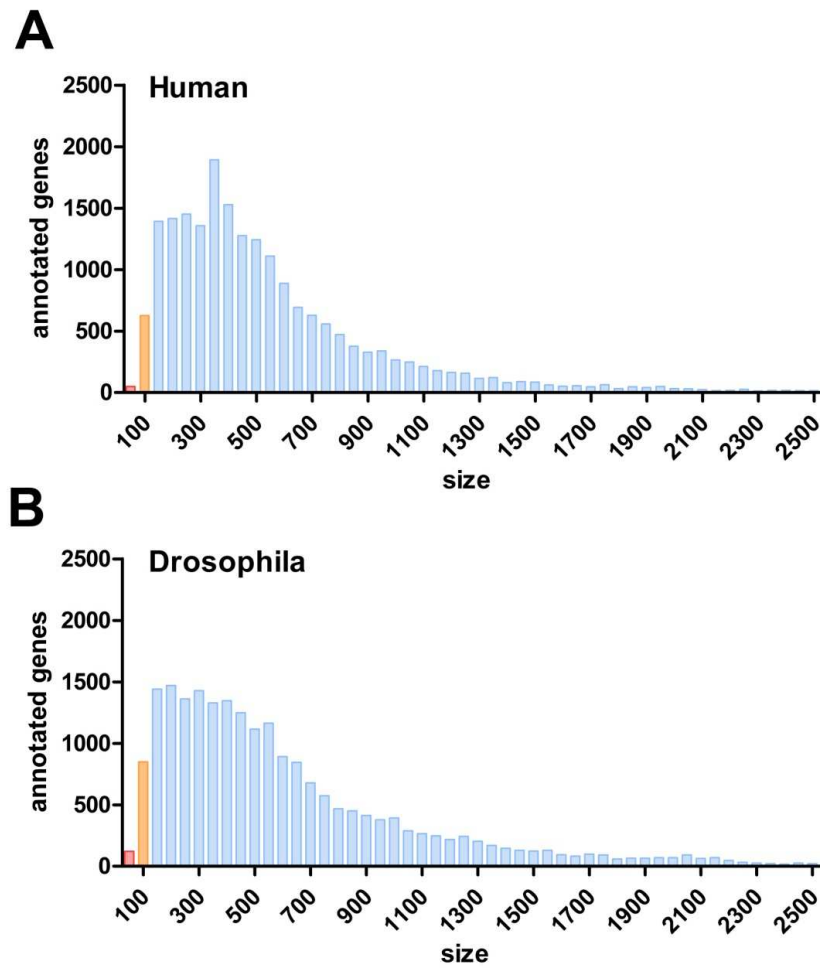


Figure 1.2

Figure 1.2: Size distribution of annotated genes in humans and *Drosophila melanogaster*.

Size distribution (in codons) of (A) Human genes as annotated in the UNIPROT database, and (B) *Drosophila melanogaster* genes as annotated in FlyBase. Genes under 50 and 100 aa are represented in red and orange respectively, genes of all others are in blue. The total number of annotated coding and putatively non-coding genes is indicated in the insets within each panel. Note the similar steep decay in the number of genes under 100 aa long as in Figure 1.1.

1.2- *tarsal-less*: The smallest protein coding gene, plays a big role in *Drosophila* development.

In order to understand the significance of short ORFs in genomes, one could start by asking the following question: What is the smallest protein a gene can encode? Although there is no real answer to this question in terms of translational mechanisms, we could look at examples of known genes, and ask what the smallest protein known to be encoded by a gene is. Our group, and others have identified an eukaryotic gene, known as *tarsal-less* (*tal*), coding for 11-33 amino acid long peptides [11,12], the smallest reported to date (Figure 1.3A) [11,12,13]. These peptides—which have been experimentally confirmed to be translated and have been functionally characterised—can act as cell-cell signals [13] and participate in different developmental processes such as the establishment of the denticle belts and trachea in embryos, and the determination and morphogenesis of the adult leg segments [11]. In these contexts, the *tal* peptides have been shown to interact with and regulate major signalling pathways. Specifically, they regulate *Notch* [14], a highly conserved signalling pathway which plays a major role in cell fate determination [15], and this regulation has been shown to be mediated by the effects of *tal* on *shaven-baby* considered to be the master regulator of trichome formation [16]. Most importantly, homologues of *tal* have been identified throughout arthropods (Figure 1.3B), showing that this gene is not a rare occurrence in fruit flies, but a *bona fide* evolutionarily conserved family of genes.

Apart from *tal*, only a handful of *Drosophila* genes (10 in total) are annotated as coding for proteins less than 30 aa long (see appendix 1). They include the ribosomal protein RpL41 of 25 aa—which is conserved in humans—, an accessory gland protein, Acp98, also 25 aa long, and 5 other predicted protein-coding genes with unknown biological functions.

The example of *tal* proves that genes can code for peptides as small as 11 amino acids. It also shows—along, to some extent, with the few other examples of small protein coding genes—that such small peptides can play essential roles in a variety of biological processes. Uncharacterised smORFs may therefore represent an important part of our genomes.

A

gcgcagcagcaacattcgacgagtgatccatccaccgataaaaagaaaaaacagctctagacatcagaaaagctccgcagatattcagctaaagccgttaaaagtatttcctgtcgggtccgcagcaaacattgaacatttataacacaaataaacagatttctgtagtcgacttttgaacgcgcagaaaattcccaaacacacacacaaacagctgactgtattatcgccccaacaccccaacacactggtggtgataataaaagactatacaaacacagcgcgcagcaaacagcagttataaaagcttcaatccgcgctgattcaaatataacacaaaggagatcgacagcagcagcagcagcagcaaaaagccagctcggtttgtcattcaagttatttttgggtcaatatacagcgcatatcgaattggcagcctacttggatcccactggccagtactaaagaagctacacgcgcagacgaagacatcgtaattcgtagacctcttttag

M A A Y L D P T G Q Y *

1A

aaaaatccaataataatcacagatcttgcgcatggcgcgcctactggatccactggtcagttcgtcagttggaagttggagcaagcaagcagaagcagcaatattttga

M A A Y L D P T G Q Y *

2A

gttccaagccgaaaagtattttaaaccagtatcaaaatgtcgcacgatttggaccccactggcactactgaagttctctatcgcaagaactccacatagccca

M S H D L D P T G T Y *

3A

agcattctaaaggctgaataactatacccacttcaaaagctccacaaatacaatctctaaaaagctggtgatcccactggaacatccggcgccacacgcgcaga

M L D P T G T Y R R P R D

AA

gcgcaggactccgcccaaaagaggcgacaggactgcttggatccaaccgggcagttactagacgctgatatcccaacaacagtgccccataacgcccggtgcc

T Q D S R C T G K R R Q D C L D P T G T G Q Y *

B

ttatccacaaactctggkactgattggggggcgcgctggttgcgcgtctgtgcgcgcgaggagacttccagctgcgcgcgggagaagaagctggggatc

M I G G A R W L R V R G R E E T S S C R R R R R K L G I

ggggcttccccaaagcgatcttggggagccctgcgatggagactttgtattatgtagtttttgcgtagcctatcaataacctattatattaattatttat

G A S P S D L G E P C D G D F C I Y V V F A *

tattatcatactattttaaataactctgttctgctgttcacaaaacccgatacgcacacatcatatcttatacttattgttatcacacatacaca

ccatatatgttatatatatactatactgttattgctcttccaattgaaaaagattacgcgaagagattatgtttagtgctcatatttcccgagc

aatcatcggtttgtttaaatactcatttttttattgccaaagatttgaatgttcttttttctctctcgctgagagcaaggaaaaccattcgagag

cgagaaaatttggttagatcataagcgtttttaaagctatttataatgtctacacctcgaccgacatccagagaacccccaacacacacctctcacacccat

ttaatatataattataaaagaaaacccattttaaactgaaaaaataaaaaa

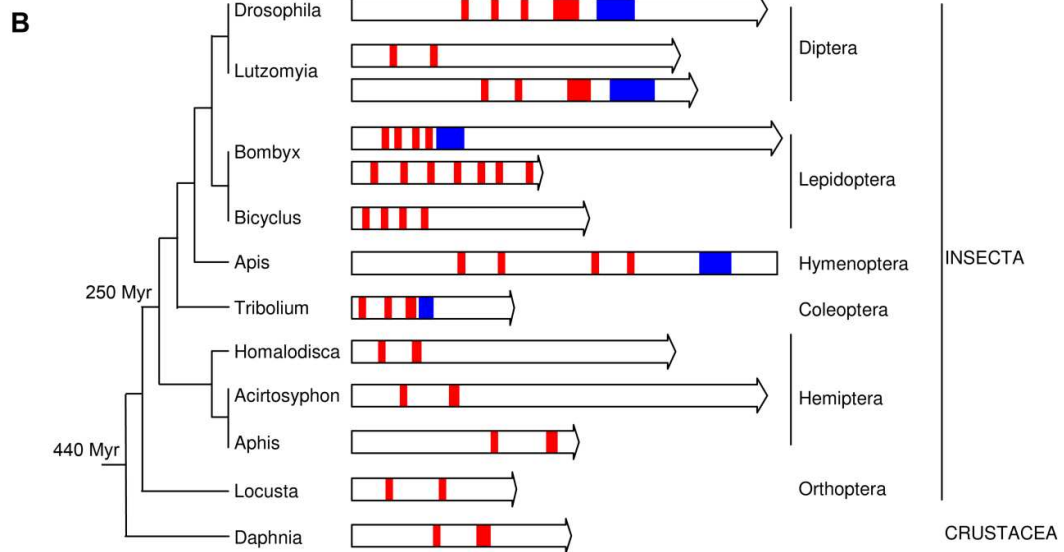


Figure 1.3:

Figure 1.3 tarsal-less, a smORF gene conserved across arthropods encodes four short peptides of 11-33 amino acids. (A) cDNA sequence of the *tal* clone LP10384.

The amino acid sequence of the translated 1A, 2A, 3A, and AA peptides are represented in red capitals underneath their respective open reading frames. Conserved amino acids between the four small peptides are in bold and Kozak sequences underlined. The B peptide in blue is not translated. (B) Graphic representation of the 440 million years (Myr) conservation of the *tal* gene family. *tal* and its homologues in other species as represented by either cDNAs (arrows) or genomic sequences (blunted arrows). This work was published in *Galindo et al. (2007)* [11].

1.3 - Targeted computational approaches predict the existence of thousands of smORFs in several organisms.

As it became evident that a pool of possible new protein coding genes with promising biological functions could be waiting to be discovered, a few attempts have been made to predict functional smORFs in the genomes of different species using targeted computational approaches. Hanada *et al.* [17] developed a “Coding Index” method based on hexamere sequence composition-statistics obtained from known coding and noncoding DNA sequences (CDSs and NCDSs respectively). In order to identify smORFs in the *Arabidopsis thaliana* genome, they combined the “Coding Index” study with analyses of conservation, transcription and purifying selection; protein coding regions, expected to be under purifying selection, should have a larger number of synonymous substitutions (Ks), which preserve the aa sequence than non-synonymous mutations (Ka), which change the aa sequence. They estimate that up to 3,241 short ORFs could be translated and belong to novel genes. A similar smORF search, was carried out by Frith *et al.* [18] in the FANTOM collection of mouse cDNAs. In that study, they used CRITICA (Coding Region Identification Tool Invoking Comparative Analysis), a gene prediction algorithm originally used for bacterial genomes, which integrates a purifying selection analysis of pair-wise aligned homologous regions into a hexamere sequence composition-analysis (similar to the one used by Hanada *et al.*) in order to increase its accuracy. They filtered their initial prediction of ~50,000 ORF candidates of all sizes for possible artefacts that may enrich the putative smORF portion, such as redundancy, sequencing errors, 5’truncation of the cDNAs, intron retention, and for short ORFs that overlap longer ORFs. They obtained 1,240 putative smORFs that could be translated, of which 495 lack similarity to any known protein. Interestingly, they compared their ORF prediction method with other commonly used “gene finders” such as Genescan, GeneID, Ensemble and Ecgene and found that although they all performed similarly well for long ORFs, the CRITICA method predicted many more smORFs, confirming the bias that the common methods have against short ORFs. What is also important to point out is that even the CRITICA method did not seem to perform very well with sequences under 50 aa, having failed to identify *sarcolipin*, a 30 aa well characterised smORF. Other considerations to take into account is that this study was performed on an extensive, although not exhaustive,

library of cDNA sequences instead of genomic DNA, and used a variety of conservative filters that may have removed *bona fide* smORFs in order to avoid possible false positive hits, which means, together with the above-mentioned 50 aa threshold, that their final estimation of 1,240 may probably be an underestimation of the number of smORFs in mice.

Our own group performed a search for unannotated smORFs in the *Drosophila melanogaster* genome [19]. We opted for a conservation based approach, selecting the ORFs that appeared to be conserved in the *Drosophila pseudoobscura* genome at the amino acid sequence level, which eliminated the noise from silent mutations at the DNA sequence level. The two fly species having diverged approximately 25 to 55 million years ago, it would be expected that putatively neutral sequences would not show any significant conservation. Conserved ORFs needed to have tBLASTn (Basic Local Alignment Search Tool using a translated nucleotide database) hits, with a stringent E value threshold of $E=1 \times 10^{-3}$, which was found to produce a false discovery rate of 7% using a set of random ORFs. Importantly, these sequences also needed to have a start and stop codon within 100 codons from the tBLASTn hit in the *D. pseudoobscura* genome. This way, from the 556,554 short ORFs initially detected in the *D. melanogaster* genome, 4,561 were considered to be conserved in the *D. pseudoobscura* genome; this is the upper estimate of the number of un-annotated smORFs in the *Drosophila* genome. We ran these sequences through a set of filters selecting for hallmarks of functionality such as synteny, which guarantees that similarity in the sequences is due to homology; evidence of purifying selection, selecting for sequences that pass a very conservative threshold of $Ka/Ks < 0.01$; and evidence of transcription. In total, 401 sequences passed all the filters; this is our conservative estimate of un-annotated functional smORFs. As with the Frith *et al.* mouse study, this number is likely to be an underestimation, since we only considered ORFs encoded by contiguous nucleotides, thereby eliminating any possible ORF interrupted by an intron. Moreover, our filters also appear to be very conservative because only 7 out of 25 annotated smORFs with proteomics evidence of translation pass them all; most of the smORFs pass the conservation and transcription filters, but several fail the synteny and stringent purifying selection filters. Interestingly, the upper and conservative pools of smORFs have size distributions with medians of <20 aa (19 and 17 aa respectively), matching the range of sizes that current genome annotations

appear to be the most devoid of (Figures 1.1 and 1.4). The size distribution of these putative smORFs is very different from that of the subset of ORFs that show sequence conservation but not start and stop codon conservation, and is also very different from that of artificially generated equivalent control sequences submitted to the same pipeline (Figure 1.4C and D), showing that those distributions (and hence the smORFs generating them) are genuine, and not randomly generated.

A similar study, in yeast [20], was performed using similar techniques to select for smORFs that are conserved across all fungi, transcribed, and with favourable codon adaptation indexes. 558 smORFs pass these criteria, and represent their estimate of putative un-annotated smORFs in yeast.

Finally, a study performed in prokaryotes [21], which also used conservation, as well as the identification of prokaryotic ribosomal binding sites known as *Shine-Dalgarno* sequences, in order to identify novel smORFs, reported that up to 2,000 small ORFs between 16-50 aa long could be translated in *E. coli*. Interestingly, it appears that the majority of these short peptides (39 out of 60 experimentally verified smORFs) have a predicted hydrophobic, single trans-membrane α -helix structure, indicating that most of those peptides may have a cell-membrane-related function. In accordance with this observation, they observed that the translation products of several of these putative smORFs, was preferentially detected in membrane rather than cytoplasmic fractions.

Overall, the four independent studies in eukaryotes predict a similar proportion of un-annotated smORFs within their respective organisms (between 5 to 25% new smORFs compared with annotated genes, depending on how conservative the thresholds are for these estimations), while the bacterial study proposes a larger estimate (almost 50% compared with annotated *E. coli* genes). All of these studies suggest that in the genomes of most species, short ORFs have indeed been under-annotated, and again, support the idea that smORFs are an abundant genomic element.

Figure 1.4: Size distribution of different pools of predicted smORFs in *Drosophila melanogaster*, as published in *Ladoukakis et al.* (2011) [19]. (A) 4,561 putative smORFs with conservation of sequence and start and stop codons in *D. pseudoobscura*, representing the upper estimate for the number of smORFs in *Drosophila*; (B) 401 smORFs with conservation of sequence and start and stop codons in *D. pseudoobscura*, with a Ka/Ks score < 0.1 , and also present in syntenic and transcribed regions, this represents the conservative estimate of smORFs in *Drosophila melanogaster*; (C) 43,197 smORFs with tBLASTn hits with E-value $< 1 \times 10^{-3}$ representing putative smORFs with some kind of sequence conservation in *D. pseudoobscura*. (D) Comparison of size distribution cumulative densities between the 4,561 putative smORFs with conservation of sequence and start and stop codons in *D. pseudoobscura* (SS), and a subset of artificial smORF like controls composed of reverse stop-to-start control ‘smORFs’ passing the same filters. The size distribution of the candidate ‘real’ smORF is significantly different from that of the controls representing random short DNA sequences.

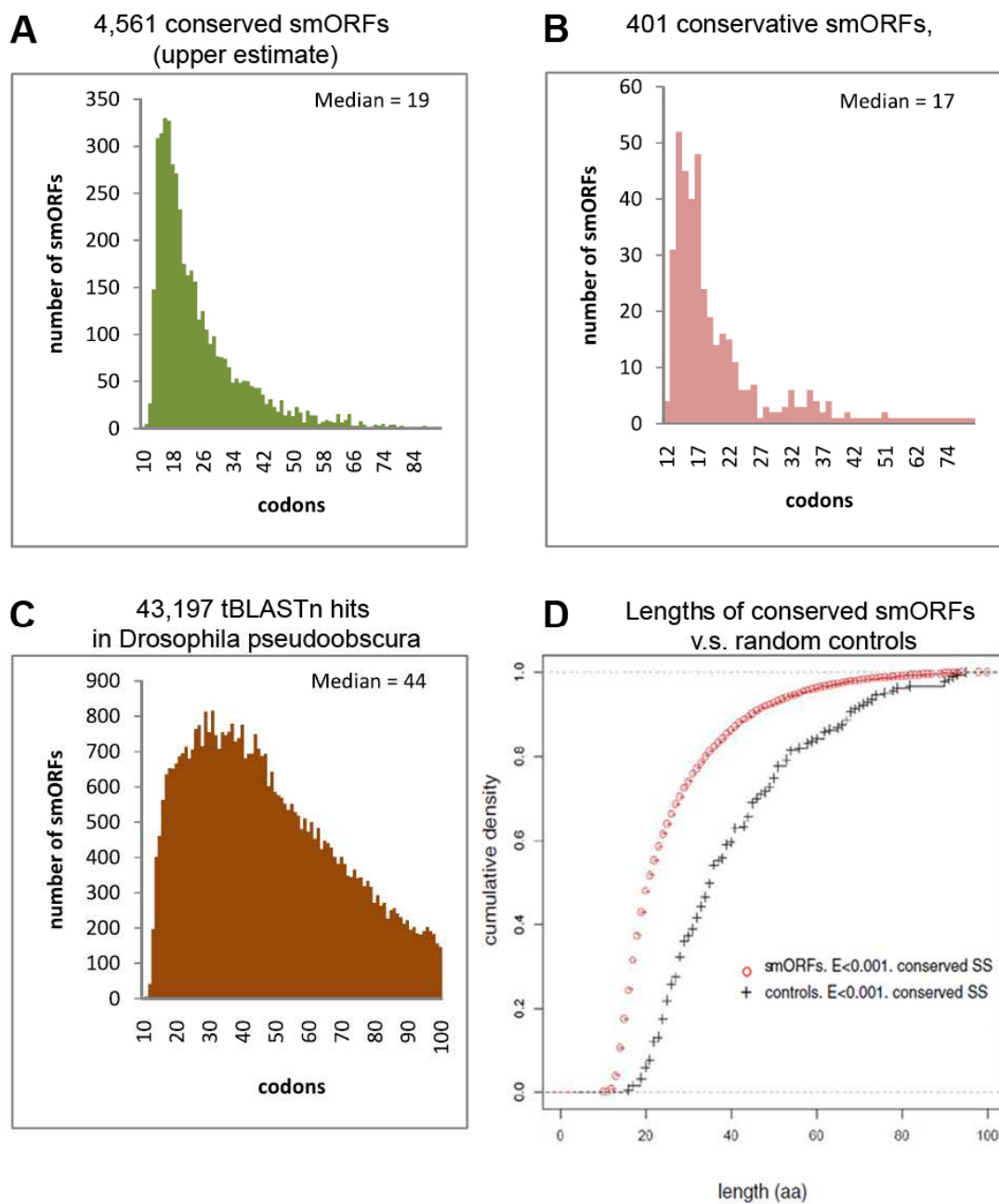


Figure 1.4

1.4 - Assessing the function of new smORFs

Even if there is extensive evidence of the existence of several smORFs in different genomes, the attribution of biological functions to these short peptide coding genes would be the most convincing proof of their relevance, yet only few studies have directly addressed their function. One of the main reasons for this lack of functional analyses, apart from the fact that smORFs are a relatively recent discovery, may be that such analyses are hard to implement on a large scale. This is particularly true in eukaryotes where the generation of mutants, even with the most advanced genetic tools available, can still be challenging. Besides, this sort of functional characterisation would need a “reverse genetics” approach, in which the effect of disrupting the expression of a particular gene is unknown. Therefore, assessing the phenotype of the affected animals or cells is not necessarily straight forward; this will be discussed further in section 2.1 of this chapter.

Some of the studies that were discussed above, which focused on the detection of smORFs, provide limited evidence of their functionality. In the Frith *et al.* mouse study [18], the potential to have a biological function was inferred for the majority of their 1,240 detected smORFs from their patterns of expression. Expression data obtained from tissue-specific micro-arrays shows that most of these mouse smORFs appear to be expressed in a highly tissue-specific manner rather than ubiquitously. Indeed the majority of those smORFs could be clustered into different groups sharing preferential expression in specific tissues, demonstrating that those smORFs may have a tissue specific function. A subset of these smORFs (25 of them) were tagged and transfected into HeLa cells; 14 of them resulted in peptide synthesis, some of them even showing a subcellular localisation, which correlated with their predicted secondary structures. These smORFs are therefore translated into peptides that may have specific cellular functions. In the Kessler *et al.* yeast study [20] the authors only examined the phenotype of a single smORF mutant, which lacked the ability to grow at 37⁰C.

Two other studies have focused primarily on the functional characterisation of smORFs, rather than on their detection, addressing the issue in a more high-throughput manner, taking advantage of the resources available for their model organisms, and scoring for specific phenotypes. In yeast, which is an organism relatively easy to manipulate genetically in order to generate mutants, Kastenmeyer *et al.* [22] generated a collection of strains carrying deletions for 140 smORFs, which they identified through a literature

search, and screened them for growth defects; 22 of those strains showed growth defects in different conditions including haploid growth, incubation at different temperatures, or with non-fermentable carbon sources, or with agents inducing DNA damage and cell replication arrest.

In *Arabidopsis thaliana*, an organism in which the generation of transgenic plants is relatively easily achieved, due to its susceptibility to horizontal gene transfer by *Agrobacterium tumefaciens* upon infection, Hanada *et al.* assessed the functionality of their previously predicted smORFs, by over-expressing several hundreds of them and screening for morphological defects in the plants [23]. For this screen they tested 473 smORFs, deemed the most likely to be functional amongst the candidates from their previous work [17] because of their high levels of transcription, determined, again, by micro-array analysis of tissue specific mRNA extracts, and evidence of conservation. 49 of these smORFs (10%) produced morphological defects upon over expression, which is almost seven times higher than when over-expressing random canonical long ORFs (572 / 40,422, or 1.4%). However, it cannot be discounted that some of the smORF-induced phenotypes are actually due to regulatory effects of the RNA transcripts themselves, rather than to their translation products.

Altogether, these studies show that a significant portion of novel smORFs have some sort of function. However, these studies have not elucidated the molecular / cellular functions of these smORFs. In order to achieve the functional characterisation of these genes at that molecular level, a more meticulous and targeted approach is necessary.

2-Using *Drosophila melanogaster* as a system to functionally characterise a novel smORF

2.1- *Drosophila melanogaster* is a good model to study the function of a novel smORF

In this project, I focus on the functional characterisation a novel smORF in *Drosophila melanogaster*. This model organism has one of the most comprehensively annotated genomes, and yet, as mentioned above may still be incomplete with respect to small open reading frame genes. The case of *tal* was extremely important in highlighting the importance of smORFs, with some of the bioinformatics studies mentioned above citing

this example to justify their search for smORFs. However, the case of *tal* is almost unique, in the sense that very few other studies have identified the specific functions conveyed by a novel smORF gene. Therefore, the functional characterisation of another gene similar to *tal* could be considered as a case study that would reinforce the importance of smORFs, while contributing to our general knowledge of this class of genes.

Drosophila melanogaster is a good model for this specific reverse genetics approach. The fruit fly, has been one of the most widely used models for genetic analyses for almost a century, and therefore a myriad of tools have been developed in this organism that make it possible to generate mutants lacking specific genes, or transgenic flies expressing genes in a tissue-specific manner. All of these genetic manipulations can be achieved in a relatively short time and at a reasonable cost —two important considerations for a project like this, where the outcome is not necessarily guaranteed, given that nothing is known about the gene to be studied, and in the worst-case scenario a reconsideration of the studied gene may be necessary. Most importantly, the fly, unlike other genetic models such as yeast, is a multicellular organism which has a variety of different tissues and behaviours. It should therefore be possible to narrow a phenotypical study, focusing on the tissues that show expression of the gene, while allowing for a range of different kinds of phenotypes to be observed.

As mentioned in section 1.4 of this chapter, such a phenotypical study is not necessarily straightforward. In the best case scenario the gene would have a morphological function, in which case the organism could present visible defects in the tissues where the gene is expressed, although these could still be subtle enough to be missed. The gene could also have an essential cellular function, in which case the phenotype would be the lethality, or poor viability of the mutant organism, or an atrophy or absence of the tissues expressing the gene. On the other hand the gene could have a physiological or metabolic function, in which case the phenotype could be much harder to assess, since it may lead to a dysfunction that is not apparent unless a very specific method is employed to detect it. Similarly, the mutant organism could also present a behavioural rather than morphological phenotype. Finally, and in reference to the above-mentioned worst case scenario, it is possible that for a variety of reasons, like redundancy for example, a smORF mutant may simply generate no phenotype at all. This variety of phenotypes and outcomes is the reason why it is difficult to implement a high-throughput functional

screen for a large number of putative genes. A study focused on a single gene, on the other hand, could take advantage of all these possible outcomes to attribute a particular function to that gene; one could address the issue in some sort of process of elimination, scoring for the most obvious phenotypes first, and moving to the more subtle ones, until a phenotype is detected.

2.2-The putative noncoding RNA gene *pncr003;2L* is a very interesting candidate for this study

In order to choose a good putative smORF candidate for this study, the best approach would be to consider a gene that fulfils all, or at least most of the parameters that have so far been linked with functionality, while being in a genomic position that favours its genetic characterisation. Such a gene was identified amongst the list of genes described as noncoding by Tupy *et al.*[24], the study in which *tal* and *RpL41* were erroneously deemed as non-coding.

This putative non-coding RNA, annotated as *pncr003;2L* (for putative non coding RNA 003 in 2L) encodes for a short ORF of 28 amino acids, which has very similar scores as those of one of the *tal* peptides of similar size (*tal* AA, of 33 aa), when subjected to the conservative filters from the above-mentioned smORF detection pipeline in *Drosophila* [19] (Table 1.1). This short ORF, therefore, appears to be evolutionarily conserved, has undergone purifying selection, and has strong evidence of transcription, which altogether support its translation. Furthermore, *pncr003;2L*, along with its *Drosophila pseudoobscura* orthologue, appears to be expressed in the embryonic somatic muscles [24]. This is interesting since tissue-specific expression has also been associated with functionality [18], but most importantly, while providing a context for focusing the phenotypical study of this putative smORF coding gene.

The *pncr003;2L* gene is located in the left arm of the autosomal chromosome II, in a locus which does not overlap any other annotated gene, and most importantly, there is a transgenic line publically available from the Harvard Medical School Exelixis stock centre, which carries a transposon inserted in the putative 3'UTR of *pncr003;2L*. The implications of this insertion will be discussed in more detail in [Chapter IV](#), where I will address different strategies to disrupt the expression of *pncr003;2L*, using this transposon.

Finally, being considered and currently annotated as a non-coding RNA, *pncr003;2L* would be a particularly interesting candidate for a case study; if found to be protein-coding, it would provide yet another example, along with *tal*, that non-coding genes may sometimes be misannotated.

| smORF | codons | translation | tBLASTtn | smORF in <i>Dp</i> | transcription | Ka/Ks |
|-------------------|--------|-------------|----------|--------------------|---------------|-------|
| <i>tal</i> AA | 33 | Yes | 4.00E-11 | yes | yes | 0.03 |
| <i>pncr003;2L</i> | 28 | not known | 8.00E-10 | yes | yes | 0.19 |

Table Key: *Dp*: *Drosophila pseudoobscura*. *Ka/Ks*: non-synonymous mutations / synonymous substitutions

Table 1.1

Table 1.1: The conservation scores of a small ORF encoded by *pncr003;2L* are very similar to those of translated *tal* AA peptide: A small open reading frame within the putatively non-coding gene *pncr003;2L* has strong transcriptional evidence, and is conserved in *Drosophila pseudoobscura* (*Dp*), with similar tBLASTn values, and Ka/Ks scores as the *tal* AA peptide, which has been proven to be translated [11,25].

2.3-Objectives

2.3.a- Characterisation of the *pncr003;2L* transcript, and its encoded smORFs.

The first part of this thesis focuses on the characterisation of the transcript expression and translation of *pncr003;2L*, and its aims are: 1) To provide an initial gene model for this gene based on the information currently available. 2) To assess its transcriptional expression throughout the life cycle of the fly, and 3) to test whether this small open reading frame-coding gene is translated. For this purpose, the expression of the *pncr003;2L* transcript is assessed using classical mRNA detection techniques such as *in situ* hybridisation and Reverse Transcription-Polymerase Chain Reaction (RT-PCR), and its translation is tested by the generation of specific smORF-GFP (Green Fluorescent Protein) fusion constructs, which preserved the original context of translation of the small ORF.

This work, covered in Chapter III, shows that *pncr003;2L* is expressed in embryonic somatic muscles, and provides evidence for the prevalent expression of this gene in muscle tissues throughout the life cycle of *Drosophila*, including cardiomyocytes in the larval and adult stages. Interestingly, it is shown that this gene encodes for two related small ORFs, both of which are shown to be translated, with their peptides having a membrane-like subcellular localisation. When expressed in the muscle tissues of transgenic flies, the peptides localise to the dyads; the structures where the T-tubules contact the sarcoendoplasmic reticulum (SER), and which ultimately control muscle contraction and relaxation through the release and uptake of calcium by the SER, which suggests that these peptides may have a physiological role during muscle contraction.

2.3.b-Characterisation of the effects of the *pBac{WH}fF02056* insertion, and its use to generate *pncr003;2L* null alleles.

The second part of this thesis focuses on the use of the *pBac{WH}fF02056* transposable element as a tool to disrupt the expression of *pncr003;2L*, in order to characterise the function of this gene. The existence of this particular insertion was a determinant factor for the choice of the *pncr003;2L* putative smORF for this study, as it allows for the implementation of relatively well established genetics methods to generate null mutants. The aim of the work presented in Chapter IV, is firstly, to assess the effects of this insertion on the *pncr003;2L* smORF gene, and secondly, to use this transposable

element in two different mutagenesis strategies to obtain a null condition for *pncr003;2L*.

The first part of this work focuses on the identification of an interesting muscle phenotype associated with the *pBac{WH}F02056* insertion, which happens to be similar, and additive to that of the major muscle gene *Myosin heavy chain (Mhc)*. Although initially very interesting with regards to a possible function for the smORF gene, the work throughout this chapter shows that this phenotype is independent of *pncr003;2L* itself, and is caused by a background mutation linked to the *pBac{WH}F02056* insertion. The background mutation was mapped, using a combination of genetic and molecular methods, to the *Mhc* locus itself, which explains the phenotypical observations.

In the second part of this chapter, the *pBac{WH}F02056* insertion was successfully used in two different mutagenesis methods, involving the generation of a specific small genomic deletion by taking advantage of the FRT recombination site within the *pBac{WH}F02056* element, and the generation of genomic lesions via the use of ionising irradiation, which although non-specific, were screened to isolate a condition affecting the *pncr003;2L* locus. The combination of the resulting mutants from these two methods led to the generation of a *pncr003;2L* null genotype. Although the *pncr003;2L* null flies show no visible phenotype, which would be in accordance with the *pncr003;2L* gene having a subtle physiological function, this null condition can be used in more specialised muscle function assays to identify the function of the smORF gene.

2.3.c-Functional assessment of *pncr003;2L* in a specific physiological context:

The aim of the third part of this thesis is to use the information and tools gathered throughout this work, including the *pncr003;2L* null genotype, to attribute a function to the peptides encoded by this gene, within a specific physiological context. In Chapter V, an extensive analysis is presented, of the effects of *pncr003;2L* in the contracting heart, which is a system that has been proven to provide a sensitive, yet relatively simple, way to assess muscle contraction in *Drosophila melanogaster* [26,27,28,29]. The study presented in this chapter focuses on the effects of the peptides encoded by *pncr003;2L* in heart contraction, and the calcium dynamics underlying this process. In this study it is shown that lack or excess of function of *pncr003;2L* results in heart

arrhythmias and abnormal calcium transients during heart contraction, suggesting that, in accordance with its subcellular localisation, *pncr003;2L* has a physiological role in the regulation of calcium cycling during muscle contraction.

2.3.d- Identification of the molecular context of the *pncr003;2L* function through an extended homology search.

The aim of the final part of this thesis, is to determine the specific molecular process by which the *pncr003;2L* peptides exert their function. For this, I explored the potential of a powerful protein homology search engine, named PHYRE2 (protein homology/analogy recognition engine) [30], in order to identify possible homologous sequences for *pncr003;2L*, which may shed light onto the molecular function of its encoded peptides. This homology search method, which has been shown to be successful in identifying remote homologous sequences by searching for hits with structural as well as sequence similarity, identified *Sarcolipin* (*sln*), a human smORF known to regulate calcium function in muscles through the inhibition of the sarcoendoplasmic reticulum Ca^{2+} ATPase (*SERCA*), as a possible homologue for the *pncr003;2L* ORFs. The work presented in this chapter studies the functional relationship between the *Drosophila* *SERCA* homologue (*Ca-P60a*) and *pncr003;2L* and ultimately supports the homology between the vertebrate *sarcolipin* / *phospholamban* (*pln*) family of *SERCA* inhibitors and *pncr003;2L*, showing that *pncr003;2L* belongs to a highly conserved family of smORFs which constitute an ancient regulatory system of cardiac function, and muscle contraction.

Chapter II - Materials and methods.

In this chapter I describe the materials and methods used for the elaboration of this thesis.

2.1-Fly strains.

Fly stocks and crosses were cultured at 25°C, in plastic tubes on a modified Lewis medium containing yeast, agar, cornmeal and glucose with Nipagen [31]. The Oregon-Red (Or-R) line was used as wild-type strain. The following lines used in this work were obtained from the Bloomington Stock Centre (Listed in the order their order of appearance):

yw;;Dmef2-GaL4, w;Dmef2-GaL4,UAS-mcd8RFP/Kr; w;DfEd1153/CyO, w;DfExcel 7067/ CyO, w;DfExcel 8036 /CyO, yw;P{lacW}Mhc^{k10423}/CyO, w;Δ2-3Sb b /Tm6b, chif^d cn¹ sca¹ bw¹ sp¹/CyO, P{ry=hsFLP}1,w; Adv /CyO, w; b cn bw, w;Gla BC/ CyO, w;DfED1102/ CyO, w;DfBSC325/ CyO, w;DfED1109 / CyO, w;DfED2256 / CyO, wPBac{ RB}CG42389^{e02963}/CyO, yw; Mi{MIC}Apep^{MI01970}/SM6a w;Mi{ET1}Cas^{MB08748}mdy^{MB08748}/SM6a, w;;UAS-GCaMP3, w; Ca-P60A^{Kum295}/CyO,

The *w; tin-GaL4* line was a gift from Manfred Frasch from the University of Erlangen-Nuremberg. The *pBac{RB} e01605* and the *pBac{WH} F02056* lines were obtained from the Harvard Exelixis collection.

2.2-Gamma ray mutagenesis.

Gamma-ray mutagenesis was used to generate another deficiency (*Df(2L)scl^{g6}*) spanning at least the 80 Kb genomic region between the genes *CG13282* and *Apep^P*, with an undetermined breakpoint somewhere in the 300 Kb region between *Apep^P* and *CG31784*. Two to seven day old males, homozygous for the *white* mini-gene-bearing insertion *pBac{WH} F02056* mapping to the 3' UTR of *pncr003:2L*, were irradiated with 4500 rads of gamma-rays (using a ⁶⁰Co source) and crossed to *yw; CyO / Gla, Bc* females; the progeny was then screened for loss of the *white* marker, and stable stocks were generated from each individual *white* eyed mutant. The span of these deficiencies was determined with the appropriate genetic complementation crosses and PCRs (Chapter IV, Figure 4.13).

2.3-RNA-extraction.

RNA was isolated using the TRIzol reagent (Invitrogen). The tissues were frozen in dry ice, and homogenised using a sterile pestle in a 1.5 mL eppendorf tube containing 200 μ L of TRIzol reagent, after homogenisation, 500 μ L more of TRIzol were added. After addition of 200 μ L of chloroform:isoamyl alcohol (24:1) the samples were incubated for 5 minutes at room temperature, spun to separate phases (12.00 x g, 4°C), and the resulting supernatant was transferred to a sterile, nuclease-free tube. The RNA was then precipitated with 500 μ L of isopropanol by centrifugation (10.00 x g, 4°C), and pellet washed with 75% ethanol before being re-suspended in 50 μ L of nuclease-free H₂O. RNA samples were stored at -80°C.

2.4-DNA-extraction.

Genomic DNA was isolated from 20-25 flies using the Wizard genomic DNA purification kit (Promega) following the instructions of the manufacturer. The flies were frozen in dry ice and homogenised, using a sterile pestle, in an eppendorf with 600 μ L of chilled Nuclei Lysis Solution. The samples were incubated at 65°C for 30 minutes, followed by the addition of 3 μ L of RNase Solution and then incubated at 37°C for 30 more minutes. 200 μ L of Protein Precipitation reagent were then added, and the samples vortexed and incubated on ice for 5 min. Proteins were precipitated by centrifugation at 16,000 x g for 4 min. The supernatant was then transferred to a fresh eppendorf tube containing 600 μ L of isopropanol, mixed, and centrifuged at 16.000 x g for 1 min. The supernatant was discarded, and 600 μ L of 70% ethanol were added to the sample. The samples were centrifuged at 16.000 x g for 1 min and the supernatant discarded. The pellet was left to air-dry for 15 min, and rehydrated in 100 μ L of rehydration solution for 1 hour at 65°C. DNA samples were stored at -20°C.

2.5-cDNA synthesis.

cDNA was synthesised using the RETROscript kit (Ambion) following the protocol for the 'Two-step RT-PCR with heat denaturation of RNA' procedure provided by the manufacturer. 2 μ g of total RNA was combined with oligo (dT) primers (2 μ L from a 50 μ M stock solution) and nuclease-free water, then denatured at 80°C before the addition of the remaining RT reagents: 10X RT buffer, dNTPs (2 μ L from a stock solution containing 2.5 mM of each dNTP), RNase inhibitor (0.25 units), and the M-MLV Reverse Transcriptase (2.5 units). Reverse transcription of cDNA was done at 42°C for 1-2 hours. The reaction was stopped by inactivating the reverse transcriptase at

92°C for 10 minutes. Newly synthesised cDNA was stored at -80°C until ready to use in PCR reactions.

2.6-Polymerase Chain Reaction (PCR).

Standard PCR reactions were conducted using reagents provided in the Taq PCR Core Kit (QIAGEN). PCR reactions were prepared on ice to a total volume of 50 µl as follows: 5 µl 10X PCR buffer; 10 µl 5X Q-Solution; 1 µl MgCl₂ (25mM); 1 µl dNTP mix (from a stock solution containing 2.5 mM of each dNTP); 1 µl of each forward/reverse specific primer (from a stock solution 100µM); 2 µl template cDNA; 0.25 µl Taq DNA polymerase (5 Units/µl); H₂O up to 50 µl.

Long PCR reactions (yielding products longer than 3Kb), and reactions requiring a mutation free product (such as the fragments used to generate the *CG21739* and *pncr003;2L*, *pln*, *sln*, rescue, over-expression and tagged constructs, see **Plasmid construction**, in this chapter) were carried out using the Expand long template PCR system (ROCHE). PCR reactions were prepared on ice to a total volume of 50 µl as follows: 5 µl 10X PCR buffer; 1 µl DMSO; 1 µl MgCl₂ (25mM); 2.5 µl dNTP mix (from a stock solution containing 2.5 mM of each dNTP); 2.5 µl of each forward/reverse specific primer (from a stock solution 100µM) ; 2 µl template cDNA; 1 µl DNA polymerase (5 Units/µl); H₂O up to 50 µl.

PCRs were performed using either a Techne TC-3000 or an Eppendorf Mastercycler Gradient thermocycler. Cycling conditions were optimised based on specific primers, melting temperatures (T_m), and length of the expected PCR product.

Standard PCR conditions:

DNA denaturation at 94°C – 5 minutes

25-30 cycles of: - denaturation at 94°C – 30 seconds
 - annealing at 5°C below average primer T_m – 30 seconds,
 - extension at 72°C – 30 seconds to 2 minutes (depending on target length)

Extension at 72°C – 10 minutes

Hold: 4°C

For semi-quantitative RT-PCRs, the reactions were initially performed with a series of different cycles, ranging from 15-35 cycles, in intervals of 5 cycles, in order to identify the number of cycles yielding an amount of product falling within the exponential phase of the PCR reaction. This number of cycles was used for experimental quantification.

2.7 -List of primers:

FW: forward and RV: reverse

ORFA / B GFP and mCherry constructs and pncr003;2L cloning

| | |
|--------------------|---|
| mCherry HindIII FW | 5' CAAGCTTGTGAGCAAGGGCGAGGAGGATA 3' |
| mCherry HindIII RV | 5' AAGCTTAGGGCTTGTACAGCTCGTCCATGCCGC 3' |
| GFP HINDIII FV | 5' AAGCTTGTGAGCAAGGGCGAGGAGCTG 3' |
| GFP HINDIII RV | 5' AAGCTTAGGGCTTGTACAGCTCGTCCATGCCGA 3' |
| ORFA Hind-III FW | 5' AAGCTTAATGTTTCCGGCAAGTAGATGGTCCTTAGGGCAGG 3' |
| ORFA Hind-III RV | 5' AAGCTTCAATACGGCATAGATGAGGTAGAGGAAGAAAAGC 3' |
| ORFB Hind-III RV | 5' AAGCTTGAAGGCGGCTTCGTAGAAGGCATAGA 3' |
| ORFB Hind-III FW | 5' AAGCTTGCCACAGCCTCAAGTCACCCATGA 3' |
| Exon2'Cdna FW | 5' TCTCTCGAATTCTTTATTCCTGCAGTTTGTGTTGCTGTT 3' |
| Exon2'Cdna RV | 5' TCTCTCGCGGCCGCGAGTTATTGCGCGCCTTTAGCT 3' |
| ORFA FW | 5' GTGTGTGGCGGCCGCGTTGAGCCAAAGGCTTTCA 3' |
| ORFA RV | 5' CGTGTGTGGGTACCCTGCCCTAAGGACCATCTACT 3' |
| ORFB FW | 5' GTGTGTGTGAATTCTTAGGGCAGGACCAAAGCC 3' |
| ORFB RV | 5' TCTAGGCCACAGCCTCAAGGCGGCCGCACACACAC 3' |

Myosin heavy chain exon sequencing

| | |
|---------------------|--------------------------------|
| fragment FW 1 | 5' ATCCCGCAATCCCCCATAGA 3' |
| fragment RV 1 | 5' TCGGATCGTAGTTAAAGCACCACA 3' |
| fragment seq RV 1 | 5' GGTAGCAGCAGCATCAGCGG 3' |
| fragment FW 2-3 | 5' CGCTATTGCTGCTGCTGTC 3' |
| fragment RV 2-3 | 5' GTGAGTGATTGGCGGTAGATAAG 3' |
| fragment seq FW 2-3 | 5' AATAGTATGCTTTTCTGA 3' |
| fragment seq RV 2-3 | 5' AGAACATAGAACGCATACTTG 3' |
| fragment FW 4-5-6 | 5' GAGCACTCGGAAAAGTGA 3' |
| fragment RV 4-5-6 | 5' TGCCCTGGGAGACAATG 3' |

| | |
|--------------------------|---------------------------------|
| fragment seq FW 4-5-6 | 5' TCGCCCGATACAAAACTAA 3' |
| fragment seq RV 4-5-6 | 5' TGAACCTAAACCACAACCTAAAAGA 3' |
| fragment FW 12 | 5' CCCCAGAAGCTCCCAGAACAGT 3' |
| fragment RV 12 | 5' CCAATTACCCCAGACAGTGACCAA 3' |
| fragment seq FW 12 | 5' ACAGTGTTGGTTCCGCTTAGTGC 3' |
| fragment FW 13-14-15 | 5' TCCACCGAATCGACCACACC 3' |
| fragment RV 13-14-15 | 5' AAATACAGAGGCGAGAAGCGAGAG 3' |
| fragment seq FW 13-14-15 | 5' ACCGCCGTCGAACCACCAC 3' |
| fragment seq RV 13-14-15 | 5' TTTCCAAATAACCTTCAAT 3' |
| fragment FW 19 | 5' CAAAACGTGTTTCAGGGAGTGCT 3' |
| fragment RV 19 | 5' CTGGGGCGGGAAAGTAGGAC 3' |
| fragment seq FW 19 | 5' GTCGTACTCGTTATCGTTCTATCC 3' |
| fragment seq RV 19 | 5' ATGGCCAGTAAATATGAATGAA 3' |
| fragment FW 7 | 5' CACAAAGATAATGCCCAAGTCG 3' |
| fragment RV 7 | 5' GGCATAGCTCATCGGTTCTGTG 3' |
| fragment seq FW 7 | 5' CACACTGCAAACACTTCACAC 3' |
| fragment FW 8-9 | 5' TATCCGTAGCACCCGTAGG 3' |
| fragment RV 8-9 | 5' TCCGCAGATTTTCGATTACAT 3' |
| fragment seq FW 8-9 | 5' AAAAATGCTCAAAAACAAACC 3' |
| fragment seq RV 8-9 | 5' GACATGACATAACAAACGAAAATA 3' |
| fragment FW 10-11 | 5' CCACTAAAATTGTAAGGGGTAAG 3' |
| fragment RV 10-11 | 5' TCAACGTGTGGGGATTCAA 3' |
| fragment seq FW 10-11 | 5' CTAATGTGTTTTTGTAAGTCGTCT 3' |
| fragment seq RV 10-11 | 5' GAAAGATACACTAGTCATACAAT 3' |
| fragment FW 16 | 5' CTAAAACGACCCACCACCACTAAA 3' |
| fragment RV 16 | 5' CCAGCTGTTTCGCGGGCATCGTC 3' |
| fragment seq FW 16 | 5' CGAAACCAAAATGCCCACTTACA 3' |
| fragment seq RV 16 | 5' GCTGCTGCTGGTAACGCTTGATG 3' |
| fragment FW 17 | 5' AGGCCCTGCGCATGAAGAAGAAGC 3' |
| fragment RV 17 | 5' ACGCGAGCAATATGAAAGGGAAGA 3' |
| fragment seq FW 17 | 5' GGATCACGCCAACAAGGTAGGT 3' |
| fragment seq RV 17 | 5' TGGGCTTTCATATTTACTTTTT 3' |
| fragment FW 18 | 5' TAGCCCTTAAGACCCCAATGAC 3' |

| | |
|--------------------|--------------------------------|
| fragment RV 18 | 5' CGACAGCGAGACGATACGGATACT 3' |
| fragment seq FW 18 | 5' ATCAGCGCCATCTCCATTCACG 3' |
| fragment seq RV 18 | 5' GCGGGAGTGGGAGGGATGAGTT 3' |

N.B. "Seq" labelled primers are internal sequencing primers

CG31739 rescue

| | |
|---------------------|--|
| CG31739 mid NheI RV | 5' TCTCTCGCTAGCATATCGTTTTGTTTCATTACCG 3' |
| CG31739 mid NheI FW | 5' TCTCTCGCTAGCTTCATCAAATTGACTA 3' |
| CG31739 FW | 5' TCTGGGGATGGAACCTCA 3' |
| CG31739 RV | 5' AAAAAGGCTTACTATACTGAACA 3' |

Semi-quantitative RT-PCR

| | |
|----------------------|---------------------------------|
| pncr003;2L A, AB RV | 5' CTGTTCTTTGCGGTTGTTATTACAC 3' |
| pncr003;2L A, AB FW | 5' ACCTCATCTATGCCGTATTGTA 3' |
| pncr003;2L B FW | 5' TTAGCTACGAACGGTTGGAAATC 3' |
| pncr003;2L B RV | 5' CTGTTCTTTGCGGTTGTTATTACAC 3' |
| pncr003;2L exon 2 FW | 5' CCGCAACTTGTTACCCACCTT 3' |
| pncr003;2L exon 2 RV | 5' GACCATCTACTTGCCGGAACATT 3' |
| pncr003;2L exon 3 FW | 5' CTCATCCTGGCCTTCCTGCTGTT 3' |
| pncr003;2L exon 3 RV | 5' GTGGGTGGTGGTTGGTGATGGT 3' |
| MHC constitutive FW | 5' TACGAGGAGGGCCAGGAGCAGTTG 3' |
| MHC constitutive RV | 5' GCGGGCGGCATCGACCATAGC 3' |
| MHC 7d spec FW | 5' GAGATGTGCTTCCTCTCC 3' |
| MHC 11e spec RV | 5' AAGCACTTTCCGGCAGCA 3' |
| Rp49.FW | 5' CCAGTCGGATCGATATGCTAA 3' |
| Rp49.RV | 5' TCTGCATGAGCAGGACCTC 3' |

RNAi construct primers

| | |
|--------------------|-------------------------------|
| pncr003;2L RNAi RV | 5' CACCGTTGAGCCAAAGGCTTTCA 3' |
| pncr003;2L RNAi FW | 5' TAGAAGGCGGCTTCGTAGAA 3' |

specific deficiency mapping

| | |
|-------------------|-------------------------------|
| fragment 1 / a FW | 5' TCCAAGCGGCGACTGAGATG 3' |
| fragment 1 / a RV | 5' TGCCCAAGCCAAAACAGAC 3' |
| fragment 2 FW | 5' TCCAAGCGGCGACTGAGATG 3' |
| fragment 2 RV | 5' CCGGCCTTGGCGTCATCT 3' |
| fragment 3 /b FW | 5' TCCAAGCGGCGACTGAGATG 3' |
| fragment 3 /b RV | 5' TTTGTCGGTCAGTAGTTTCGCGC 3' |

gamma ray deficiency mapping

| | |
|------------|--------------------------------|
| CG13282 FW | 5' GCCAGATTCGTGAAGCGTTCGG 3' |
| CG13282 RW | 5' TAAATGAAATTACATACATCAT 3' |
| CG31739 FW | 5' GGCCAGAGCAAGAAGGACTG 3' |
| CG31739 RW | 5' ATTTGTA ACTATGAATATTTAAA 3' |

2.8-*Mhc* exon sequencing.

To sequence the 12 different genomic regions corresponding to the exons constituting the *Mhc-RK* IFM specific isoform, each of the fragments was amplified by a standard PCR reaction, using as template whole fly genomic DNA from either *Or-R* or *w;pBac{WH}F02056* flies. After confirming that each reaction yielded a product of the expected size by gel electrophoresis, each PCR product was sequenced by the Eurofins company. The sequences from the *Or-R* and *w;pBac{WH}F02056* strains were aligned with the reference sequence (as annotated in FlyBase) using the SEQ-Man software from the DNASTAR suite, and the discrepancies annotated as shown in Figure 4.8.

2.9-Agarose gel electrophoresis.

Visualisation of the various RNA, cDNA, and PCR products was done using standard agarose gel electrophoresis. 0.5-1.5% agarose gels were used, according to the expected product size, with 1X TBE (89 mM Tris, 89 mM boric acid, 2 mM EDTA) and 0.5 µg/ml of ethidium bromide (Sigma). was added to the liquid agarose before pouring into the gel cast. DNA or RNA was combined with MassRuler™ DNA loading dye (Fermentas) at a proportion of 1 µl dye for every 5 µl of sample, and loaded alongside the MassRuler™ DNA Ladder Mix (80-10,000 bp fragments; Fermentas). Gel pictures

were taken using an Uvidoc gel documentation system (Uvitec Cambridge) and UviPhotoMW image analysis software.

2.10-Miniprep of plasmid DNA.

Plasmid DNA was isolated from bacterial cultures using the QIAprep Spin Miniprep Kit (QIAGEN) following the instructions provided by the manufacturer. 2 mL of overnight culture were spun to pellet the cells (3 min, 6800 X g) and supernatant discarded. Confirmation of the recovered plasmid identity was done via agarose gel electrophoresis, after digestion with appropriate restriction enzymes to confirm the cloning was successful. Plasmids were sequenced by the Eurofins company.

2.11-Plasmid construction.

Unless otherwise stated, all plasmids used in this work were obtained from the *Drosophila* Genomics Resource Center (DGRC). The UAS-*pncr003:2L* plasmid was generated by cloning the *pncr003:2L* cDNA from the RE28911 plasmid, into the pUAST-ATTB vector. To generate the pUAST-*pncr003:2L* ORFA-GFP and pUAST-*pncr003:2L* ORFB-GFP constructs, the stop codons of either of the ORFA or ORFB coding sequences were replaced with a unique HindIII restriction site by inverse PCR of the RE28911 plasmid. The EGFP coding sequence was amplified from the EGFP-C1 plasmid by PCR using HindIII primers that eliminated the start codon and allowed for the EGFP coding sequence to be cloned downstream and in frame of either *pncr003:2L* ORFA-HindIII or *pncr003:2L* ORFB-HindIII. The resulting *pncr003:2L*_ORFA-GFP and *pncr003:2L*_ORFB-GFP (ORFA-GFP and ORFB-GFP) constructs were cloned into the pUAST ATTB vector. For the UAS- *pncr003:2L*_ORFA and UAS-*pncr003:2L*_ORFB (ORFA/Sc1A and ORFB/Sc1B) constructs, a fragment comprising the ORFA or ORFB coding sequences including a small down-stream and up-stream region (86 and 16 nt for ORFA and 46 and 16 nt for ORFB), which was enough to ensure the maintenance of the Kozak sequences and optimal PCR amplification, were amplified by PCR, and cloned into the pUAST-ATTB vector. The *pncr003:2L* FS cDNA, carries insertion mutations after the start codon of both ORFs A and B generating a frameshift, which gives rise to peptides of the same size but completely different aa sequences (ORFA FS: MAKPATCSPPLASWPSCFSSSTSSMHM, ORFB FS: MMRQKVCSPSSSWPSCCSCSMPSTKAS). This cDNA was synthesised by

Eurogenetec, and cloned into the pUAST-ATTB vector to generate the UAS-*pncr003:2L_FS* construct.

The method described by Kondo et al (2006) [32] was used to generate the RNAi *pncr003:2L* construct. For this, a region corresponding to nucleotides 1-330 from the RE28911 *pncr003:2L* cDNA was amplified by PCR and cloned into the pRISE vector.

The N-terminal FLAG-Hemagglutinin *pncr003:2L* ORFA and *pncr003:2L* ORFB tagged constructs, N-terminal FLAG-Hemagglutinin *sln* and *pln* tagged constructs, and the *sln_ORFA* and *pln_ORFA* constructs, were provided by Jose Ignacio Pueyo [33].

The CG31739 rescue construct was generated by amplifying a 6695 bp genomic region (2L:16825305 – 16832100, as annotated in FlyBase) including the CG31739 gene in its entirety and 1900 bp of its upstream region, using two contiguous PCR reactions (Roche Expand Long Range) of 3696 and 3000 bp each, ligated through the addition of a unique *NheI* restriction site at the 3' end of one product and 5' end of the other. Each of the 3696 and 3000 bp products were subcloned into the TOPO-TA vector (Invitrogen), sequenced, and cloned sequentially into the pCaSpeR 5 vector, which was used for the generation of transgenic flies.

All transgenic lines, except those indicated below, were generated at Bestgene, by PhiC31 integrase-mediated site-specific transgenesis into the third chromosome using the 24749 (86Fb) receiver strain. RNAi *pncr003:2L*, N-terminal tagged lines and CG31739 genomic rescue were generated at Bestgene, by conventional random P-element transgenesis.

2.12-Adult fly, and larvae dissections.

For beating-heart video recordings and Ca^{2+} transient measurements, female flies were collected within 24 hours after eclosion and reared for 8-12 days (or 25-30 days for ageing experiments) at 25°C. Semi-intact heart preparations were performed as described by Ocorr *et al.* [29]. Adult flies were anesthetised with Flynap (Carolina Biological Supply Company) (3-5 minutes), and immobilised dorsal side down on a small petri dish with a thin coating of petroleum jelly. The head and lower half of the thorax including the ventral thoracic ganglion and legs were cut off using fine iridectomy scissors. The samples were bathed in a freshly prepared artificial hemolymph solution (108 mM Na^+ , 5 mM K^+ , 2 mM Ca^{2+} , 8 mM MgCl_2 , 1 mM

NaH₂PO₄, 4 mM NaHCO₃, 10 mM sucrose, 5 mM trehalose, and 5 mM HEPES) adjusted to pH 7.1 and oxygenated for 15 minutes prior to the dissection. The ventral cuticle was removed, and all internal organs were carefully pulled out, leaving only the intact heart and abdominal fat bodies. Fat body was cleared away using a fine capilar-based liposuction method. The samples were allowed to stabilise for 20 minutes with oxygenation before imaging. For adult indirect flight muscle immunohistochemistry, thoraces were frozen in dry ice, and bisected through the midline using a scalpel blade, and immediately fixed in 4% paraformaldehyde for 20 minutes before staining as described in section 2.19. For larval flat preparations, wandering third instar larvae were anaesthetised by submerging them in ice-cold 1X PBS for 5 minutes. The animals were then placed in a drop of cold 1X PBS on a stylgard plate and pinned ventral side down. A longitudinal incision across the dorsal midline was used to open the larvae; guts, CNS, nerves, and tracheae were removed and the cuticle pinned to the stylgard plate by the four corners resulting from the longitudinal incision. The samples were fixed in a drop of 4% paraformaldehyde for 20 minutes before staining as described in section 2.20.

2.13-Indirect Flight Muscle sarcomere length measurements.

To measure sarcomere lengths, confocal images taken with a 60X water immersion objective, and the sarcomere lengths were measured directly from the LSM Image browser files, using Image J. For each genotype 200 sarcomeres were measured as follows: 10 sarcomeres were measured for 5 different myofibrils and this for 4 individual thoraces, corresponding each to 4 different flies. One tailed, unpaired t-tests were performed to assess the statistical significance of the length differences, using the Graph-Pad prism 5 suite (GraphPad Software. Inc., La Jolla, CA).

2.14-Simple motility assay.

For the simple motility assay, the flies were individually captured from their vials and released on a constantly illuminated flat surface, at room temperature, using a fine suction tube, and their motility categorised in three different categories over a period of observation of 10 seconds: If the flies could take off and sustain flight, they were categorised as able to fly; if the flies could not take off and fly, but could still perform small jumps (of approximately 0.5 to 1 cm in length), they were categorised as able to jump; and if the flies were neither able to fly, nor jump, they were categorised as able to walk only.

2.15-Flight assay.

For the flight assay, 1 to 2 week old flies were dumped directly from their vials, in batches of 15-20 flies per assay, into a 2L measuring cylinder, in which a paraffin-coated plastic sheet covered the inner surface of the cylinder. The flies were dumped through a funnel taped to the cylinder, to ensure they all go in through the middle of the aperture of the cylinder, and the cylinder itself was solidly taped to the ground to ensure that no tipping of the cylinder, which may bias the results, would occur. The plastic sheet was recovered after each assay, and the position of the flies scored according to a scale generated by dividing the total height of the cylinder into ten equal parts. A score of 0 represents the bottom of the cylinder, and scores of 1 and 10, represent the lowest and higher regions of the cylinder. The total number of flies in each region, for each genotype was plotted in a horizontal bar chart. 150 to 200 flies were assayed per genotype. This assay was performed at room temperature.

2.16-Heart Video recordings and period measurements.

Video recordings of hearts were acquired using a Leica DMRB microscope with a 10X PL FLUOTAR dry objective equipped with a high-resolution Hamamatsu digital camera (model C848-05G01). Time lapse recordings of beating hearts were obtained using the Simple-PCI 6 suite at 32 frames/second. Based on the method of Ocorr *et al.* [27] to get a random sampling of the heart function from the flies, a single 20 second recording was made for each fly without previewing. Recordings were taken from 10-30 flies per genotype. Image J was used to generate the kymographs (with the plug-in created by J. Rietdorf and A. Seitz) and for heart period measurements. The quick transition between the systolic and diastolic states of the heart, visualised in the kymographs as a straight line perpendicular to the time progression axis was used to delimit each heart period. Arrhythmicity indices were calculated by normalizing the standard deviation to the mean of all period measurements for each recording, over the median of that same recording [26]. The arrhythmicity index values of each genotype was normalised over the arrhythmicity of age-matched wild-type flies. The fractional shortening was calculated as the difference in percentage between diastolic heart diameter and systolic heart diameter over the diastolic heart diameter. These diameters were measured directly from the heart video recordings using Image J. For rescue experiments, each of the transgenes (*UAS-pncr003:2L*, *UAS-pncr003:2L_FS*, *UAS-pncr003:2L_ORFA*, *UAS-pncr003:2L_ORFB*, *UAS-pncr003:2L_FH-ORFA*, *UAS-pncr003:2L_FH-ORFB*, *UAS-pln_ORFA* and *UAS-sln_ORFA*) was expressed in

muscles by *Dmef2-GaL4* in a *Df pncr003;2L* background. For excess of function experiments, each of the transgenes (*UAS-pncr003:2L_ORFA*, *UAS-pncr003:2L_ORFB*, *UAS-pln_ORF* and *UAS-sln_ORF*) was expressed in muscles in a wild type background by *Dmef2-GaL4*. *tin-GaL4* was also used in rescue experiments with the *UAS-pncr003:2L* and *UAS-pncr003:2L_FS* constructs. For the statistical analysis of these results, I used two-tailed Mann-Whitney U tests, which were performed using the Graph-Pad prism 5 suite (GraphPad Software. Inc., La Jolla, CA).

2.17-Calcium fluorescent recordings.

To visualise the Ca^{2+} transients during muscle contraction, I used the genetically encoded fluorescent Ca^{2+} sensor G-CaMP3 in the semi-intact heart preparations described above, and followed a method similar to that described in Lin et al. 2011 [34]. Fluorescence measurements were made using a Zeiss laser scanning microscope LSM 510 on a Zeiss Axioskop 2 microscope, with a 10X Achroplan objective. For each beating heart preparation, a ten second time-lapse recording was taken over a strip of 512/45 pixels, covering the heart along its antero-posterior axis. The time series were taken at a speed of 16.6 frames / second. Using the LSM imaging suite (version 3.2), average fluorescence intensity was recorded for each time-lapse over an area of interest of 40/30 pixels, corresponding to the maximum width of the heart at the level of the base of the conical chamber at its most contracted state. In order to obtain average values of intensity / time for all animals, and because every heart has a unique beating frequency, I normalised the time scale for each calcium signal peak, in relation to the maximum value ($t_{0\%}$) and the lowest, basal value ($t_{100\%}$). A 2nd order polynomial curve was fitted, using GraphPad Prism 5, in order to obtain comparable data points for all samples. The intensity values were normalised to the most basal value of the recording, and all values at each time point were averaged for each genotype. I used *Dmef2-GaL4* to drive the expression of the *UAS-G-CaMP3* construct in muscles in the *Df pncr003;2L* background and compared the calcium transients of *Df pncr003;2L* mutants with that of their *pncr003:2L* heterozygous siblings (wild-type control). For rescue experiments, each of the rescue constructs (*UAS-pncr003:2L*, *UAS-pln_ORFA* and *UAS-sln_ORFA*) was driven in muscles together with *UAS-GCaMP3* in a *Df pncr003;2L* background. For ectopic expression experiments, each of the ectopic expression constructs (*UAS-pncr003:2L_ORFA*, *UAS-pncr003:2L_ORFB*, *UAS-pln_ORF* and *UAS-sln_ORF*) was driven in muscles together with *UAS-G-CaMP3* in a

wild-type background. To take into consideration the buffering effect of the 10XUAS enhancer region within the rescue and ectopic expression constructs on the available pool of *GaL4* molecules, and therefore on the *GaL4* dependent fluorescence signal, I considered the ratios of fluorescence signal of the rescue and ectopic conditions over that of control animals where the non-functional *UAS-pncr003;2L FS* (expressing the *pncr003;2L* cDNA with frame-shifts in both ORFs A and B) was driven by *Dmef2-GaL4*, in the same *Df pncr003;2L* background as for the rescue experiments, or wild-type background for ectopic experiments. To render directly comparable the differences of calcium intensity across experiments done with different levels of UAS enhancers / *GaL4* molecules, I plotted the ratios of the rescue conditions over the *Df pncr003;2L*, *Dmef2>pncr003;2L_FS* control, relative to the average maximum amplitude of *Df pncr003;2L* (equivalent to the rescue experiment control with no UAS enhancers), and the ratios of the ectopic expression conditions over the *Dmef2>pncr003;2L_FS* control relative to the average maximum amplitude of the wild-type controls (equivalent to the ectopic experiment control with no UAS enhancers). All the values were then expressed as a percentage relative to the wild-type control.

2.18-Intracellular action potential recordings.

Intracellular recordings were performed, by Jeremy Niven, on the semi-intact heart preparations described above, with the addition of 120 μ M Cytochalasin D (Sigma Aldrich) to the artificial haemolymph immediately prior to the recordings to reduce heart movement and allow for more stable recordings. Intracellular recordings were made from single cardiac myocytes in the anterior heart using electrodes pulled from 10 cm borosilicate glass capillaries (1.0 mm outer diameter, 0.58 mm inner diameter; GC100F-10, Harvard Apparatus, <http://www.harvardapparatus.co.uk>) using a Sutter P97 puller (Sutter Instruments, <http://www.sutter.com>). These electrodes were filled with 2M potassium acetate to give typical resistances of 150-200 MOhms. All recordings were made using an NPI SEC-05X amplifier (NPI Electronic, <http://www.npielectronic.com>).

Throughout recordings the temperature of the flies was maintained between 22°C and 24°C. Cardiac myocytes were identified by an approximately 50-60 mV drop in membrane potential and large (>45 mV) action potentials. Intracellular recordings with action potentials (APs) smaller than 45 mV were not included in the analysis. Spontaneous cardiomyocyte action potentials were recorded in bridge

mode. Intracellular recordings were digitised at 5 kHz using a CED micro1401 A/D conversion interface and Spike 2 software (Cambridge Electronic Design, <http://www.ced.co.uk>). The height of the APs (in mV from the resting potential) was analysed offline using custom-built software and verified by hand. Action potentials were classified as single or double peaks and their proportions assessed offline by hand. Double APs were characterised as two voltage peaks of full amplitude in which the intervening voltage did not fall below 30% of the total amplitude.

2.19-*In situ* hybridisation.

DIG-labelled RNA probes were generated from a 311 bp fragment corresponding to the third exon of pncr003:2L (nucleotides 212-523 of cDNA RE28911) following manufacturer procedures (Roche). The tissues were prepared in RNase free conditions as follows: *Drosophila* embryos were dechorionated in bleach, devitellinised and fixed in a 1:1 heptane / paraformaldehyde mix for 20 minutes, and after several methanol washes, the embryos were then stored in methanol at -20°C. Adult and larval tissues were prepared as described above in section 2.12. Embryos stored in methanol were allowed to warm up to room temperature, then rehydrated in a decreasing series of ethanol dilutions (90%, 70%, 50%, 30%), and washed several times with 1X PBT (1X PBS + 0.2% Tween-20). Larval and adult tissues were washed several times with 1X PBT (1X PBS + 0.2% Tween-20) to remove excess paraformaldehyde. The rehydrated embryos or adult and larval tissues were permeabilised using Proteinase K (3 µg/ml in PBT) (Roche) for 1 hour on ice. The reaction was stopped by washing two times with 2 mg/ml glycine in PBT followed by several washes with 1X PBT. The embryos / tissues were incubated with 0.2 M HCl for 10 minutes to remove endogenous alkaline phosphatase activity, washed several times in 1X PBT, and then post-fixed in 4% paraformaldehyde for 20 minutes. The embryos / tissues were prepared for hybridisation by first washing with a 1:1 solution of PBT:Hybridisation Solution (HS – a mixture of deionised formamide, 20X standard saline citrate, heparin (50 µg/ml), Tween-20 (0.1%), boiled salmon sperm (10 mg/ml), tRNA (0.5 ngr / mL), and nuclease free H₂O) then washed with neat HS. The embryos / tissues were then pre-hybridised in HS for two or more hours in a 56°C water bath. 200-500 ng of DIG-labelled probe in HS was heated to 90°C for one minute to relax any secondary mRNA structures, and allowed to hybridise with the embryos / tissues at 56°C overnight. After hybridisation the samples were washed in HS (2 X 20 mins) and then slowly transferred to PBT in a

series of HS:PBT solutions (1:3, 1:1, 3:1), finally being brought down to room temperature in 1X PBT. Embryos were blocked for 2 hours in a solution of 1% bovine serum albumin (Sigma) and 5% normal horse serum (Vector Labs) in 1X PBT. After blocking, the tissues were incubated in a slow rotator at 4°C overnight in a 1:1000 dilution of α -DIG-AP (Roche) in PBT. The antibody was removed by washing the samples several times with PBT over 2 hours followed by three rinses with staining solution (5 M NaCl, 1 M Tris-HCl, 1M, pH7.5, MgCl₂, 0.2% Tween-20, and H₂O). A developing solution was made by combining staining solution with 4.5 μ l/ml NBT and 3.5 μ l/ml BCIP. Signal was allowed to develop in the dark and monitored periodically as to not overstain. The samples were then washed several times in 1X PBS before mounting on glass slides in Aqua-Poly/Mount (Polysciences) prior to imaging.

2.20-Immunofluorescence.

For immunofluorescence, the following primary antibodies were used: mouse anti-discs large (DSHB) used at 1:5 dilution, mouse anti-GFP (Roche) used at 1:500, Rabbit anti-RFP (Molecular Probes) used at 1:500, mouse anti-FLAG (Sigma) used at 1:1000, rabbit anti-GFP (Molecular Probes) used at 1:500, anti-SERCA (Ca-P60A)[35] was a gift from Mani Ramaswami and was used at 1:1000. Secondary antibodies used: anti-mouse-FITC, anti-rabbit-FITC, anti-mouse-Biotin, anti-rabbit-Biotin, streptavidin-rhodamine and streptavidin-FITC (Jackson ImmunoResearch). Samples were fixed in 4% paraformaldehyde, washed in PBS, and PBTx (0.1% Triton X-100), blocked and incubated in PBT (0.3% Triton X-100, 0.2% BSA) and mounted in Vectashield (Vector). Incubations were done overnight at 4°C for primary antibodies and for two hours at room temperature for secondary antibodies. Phalloidin-rhodamine and Phalloidin-Cy5 (Life Technologies) were used at 5:200, from 1.5 mL methanolic solution, with an incubation time of 20 minutes.

Fluorescence imaging was performed by confocal microscopy (LSM, Carl Zeiss). LSM Image Browser, ImageJ and Adobe Photoshop software were used for image processing.

2.21-Bioinformatics search for homologues.

To search for structural homologues I used the PHYRE2 engine (<http://www.sbg.bio.ic.ac.uk/phyre2/>) [30] using as an input the sequence of the *pncr003:2L* ORFA peptide with the normal modelling mode. To search for sequence homologues an initial search, restricted to the same taxonomic clade, was carried out in

ESTs deposited in NCBI (<http://www.ncbi.nlm.nih.gov/>) using tBLASTn with default settings (Blosom-32 matrix, Expected threshold of matches obtained purely by chance of 10, using compositional adjustment and low complexity region filters), or maximally relaxed parameters (PAM-30 matrix, Expected threshold of matches obtained purely by chance of 1000, removing compositional adjustment and low complexity region filters). The top 100 hits were scrutinised for belonging to a smORF of less than 100aa with start and stop codons, in the correct orientation and non-overlapping with longer ORFs. The complete smORFs passing this filter were then aligned using ClustalW to the query and already identified orthologues of the same phylum. A consensus weighted by phylogeny was then extracted from the alignment and the process was iterated, carrying out a new tBLASTn search with the consensus sequence. When no more homologues from the same taxonomic class were obtained in a given iteration, the tBLASTn search was expanded to the next higher-order clade.

2.22-Transmission Electron Microscopy.

Adult flies were dissected as described above, and were treated for transmission electron microscopy as follows: the samples were fixed in 4% formaldehyde + 1% glutaraldehyde in PBS for 4 hours at room temperature, then overnight at 4 °C. They were then post-fixed in 1% (w/v) osmium tetroxide in PBS for 4h at room temperature, before being dehydrated in an ethanol series. After 2 X 30min in propylene oxide, they were left overnight in 50:50 propylene oxide:TAAB low viscosity resin (TLV; TAAB Laboratories Ltd., Aldermaston, UK), then infiltrated with TLV resin over several days, with a few resin changes, before polymerising at 60°C for 16 hours. Thin (100nm) sections were cut and stained with 0.5% (w/v) aqueous, 0.22µm-filtered uranyl acetate for 1h and subsequently lead citrate for 15 min. Sections were examined in a Hitachi-7100 TEM at 100kV and images were acquired digitally with an axially-mounted (2K X 2K pixel) Gatan Ultrascan 1000 CCD camera (Gatan UK, Oxford, UK). N.b.: The fixation and sample preparation steps were performed by Julian Thorpe, who also assisted in the acquisition of the images

Chapter III - Characterisation of the gene sequence, transcript expression, and translation of *pncr003;2L*.

1- Introduction:

The *pncr003;2L* gene is currently annotated as a non-coding RNA, and there are no indications of what its function could be. As an initial step to the functional characterisation of this gene, the work presented here will briefly review the information available in FlyBase with regards to the transcript sequence, and structure of this gene, in order to provide an initial “gene model” for this putative smORF.

This work will then focus on the characterisation of the expression of the *pncr003;2L* gene. To characterise the function of any gene, and most particularly in this “reverse” genetics approach, it is essential to understand when and where the gene is expressed, as the patterns of expression of the gene provide valuable information which may already suggest a function for the gene, and would indicate, when and where to look for a phenotype once a mutant for this gene is available. One of the major aims of this chapter is therefore to provide a temporal and spatial landscape of the expression of the *pncr003;2L* transcripts during the life cycle of *Drosophila melanogaster*, using classical gene expression analysis tools such as *in situ* hybridisation and RT-PCR, which goes much beyond the current data presented by Tupy *et al.*[24], indicating that *pncr003;2L* is expressed in the embryonic somatic muscles.

The other objective of this work is to assess the translation of the putative *pncr003;2L* transcript, which is currently considered as non-coding. In the general introduction of this thesis, the *pncr003;2L* gene has been portrayed as having promising evidence of being protein-coding, because within its transcript, a putative smORF was detected,

which passed a series of stringent bioinformatics filters designed to distinguish potential small coding sequences (Table 1.1). The translation of this small ORF into a peptide remains, however, to be proven. In the work presented here, the translation of this smORF into a peptide will be assessed experimentally, using a specific smORF-GFP (Green Fluorescent Protein) fusion construct, which is designed to preserve the original context of translation of the small ORF, by placing the GFP tag, in frame with the small ORF, within the *pncr003;2L* transcript.

In this work I demonstrate that *pncr003;2L* is expressed, throughout the life cycle of *Drosophila*, in somatic and cardiac muscle tissues, and show that the gene encodes for but two related smORFs (ORFA and ORFB), which are both translated into peptides. I show that *pncr003;2L* is a complex gene, producing different isoforms, which can be regulated in a tissue-specific manner. Furthermore, I show that these peptides have a specific subcellular localisation to the dyads of muscle cells and to the plasma membrane and peri-nuclear structures, which may provide an insight into their function.

2- Results:

2.1- *pncr003;2L* codes for two small open reading frames with an optimal translation context

The FlyBase annotation (Release 5.22 – Release 5.52) (Figure 3.1A), as well as the *Tupy et. al* manuscript [24] indicate that the *pncr003;2L* gene, which spans 4 Kbs on chromosome 2L, is transcribed into a 1Kb transcript composed of five exons (exons 1-5). This transcript is supported by the existence of the complete cDNA clone RE28911, which is available from the *Drosophila* Genomics Resource Centre (DGRC) public depository, and has therefore been used as a template for many of the constructs and probes used here. In the FlyBase database, there is also an second, shorter transcript model for this gene, supported by a few ESTs, such as RE72983. Interestingly, the sequence corresponding to the first exon of that EST is different from the sequence from any of the exons that constitute the RE28911 cDNA and maps to a genomic region just upstream of exon 3, therefore the RE72983 EST represents a transcript produced by a different promoter, provided by an alternative exon, which, because of its position between exons 2 and 3, will be referred to as exon 2' (Figure 3.1B).

An analysis of the RE28911 cDNA sequence, shows that *pncr003;2L* codes for the putative small open reading frame of 28 aa (ORFA), which was the subject of the study presented in Table 1.1 in the general introduction of this thesis. This transcript also codes for another putative smORF of 29 aa (ORFB), located only 68 nt downstream of the ORFA stop codon, whose sequence is remarkably similar to that of ORFA (Figure 3.1C). These two short ORFs share high amino acid sequence similarity between them, implying that they may have been the subject of a local gene duplication event, giving rise to a transcript with a polycistronic configuration. It is interesting to note that *tal*, which is one of the most representative examples of functional smORFs characterised so far, is also polycistronic, and the peptides it encodes also share high amino acid sequence similarity between them (Figure 1.5A) [11], which may suggest that this kind of local gene duplication phenomenon leading to polycistronic transcripts could be common in this class of genes.

The sequences surrounding the ATG start codons of both of the short ORFs encoded by *pncr003;2L* are very similar to the consensus *Drosophila* Kozak sequence (Figure 3.1D, as computed by V. Pereira) [36]. The Kozak consensus sequence is a nucleotide sequence motif present around the ATG start codon of eukaryote ORFs, which is known to be associated with an optimal translation context [37]. The presence of these sequences could therefore be considered as further evidence that the short ORFs within *pncr003;2L* are translated into proteins.

Each of these two putative ORFs, ORF A and ORF B, are encoded in exons 2 and 3 respectively. Therefore, the transcript represented by the RE72983 EST contains exclusively ORF B (Figure 3.1B). A DNA alignment between exon 2', the alternative exon giving rise to this transcript, and exon 1 from RE28911, shows a good extent of conservation between these sequences (Figure 3.1E), a similarity which supports the gene duplication event that gave rise to the polycistronic transcript represented by RE28911.

Figure 3.1: *pncr003;2L*, contains two putative smORFs with an optimum context of translation. (A) Diagram representing the genomic locus of *pncr003;2L* on chromosomal arm 2L, as depicted in the FlyBase genome browser. Genes are indicated by blue arrows. The RE28911 cDNA and RE72883 EST are indicated underneath the *pncr003;2L* gene . (B) magnification of the RE28911 cDNA and RE72883 EST, showing the position of ORFs A and B. (C) cDNA sequence of the *pncr003;2L*, clone: RE28911. The amino acid sequence of each putative peptide is represented in red capitals underneath their respective open reading frames. Conserved amino acids between the two peptides are in bold, and Kozak sequences are underlined. (D) Kozak sequence consensus, obtained from the comparison of the nucleotides surrounding the ATG start codon of 16,884 protein coding genes annotated in FlyBase (unpublished data from *Pereira V.* (2008)). (E) DNA alignment of exons 1 and 2' showing their sequence conservation.

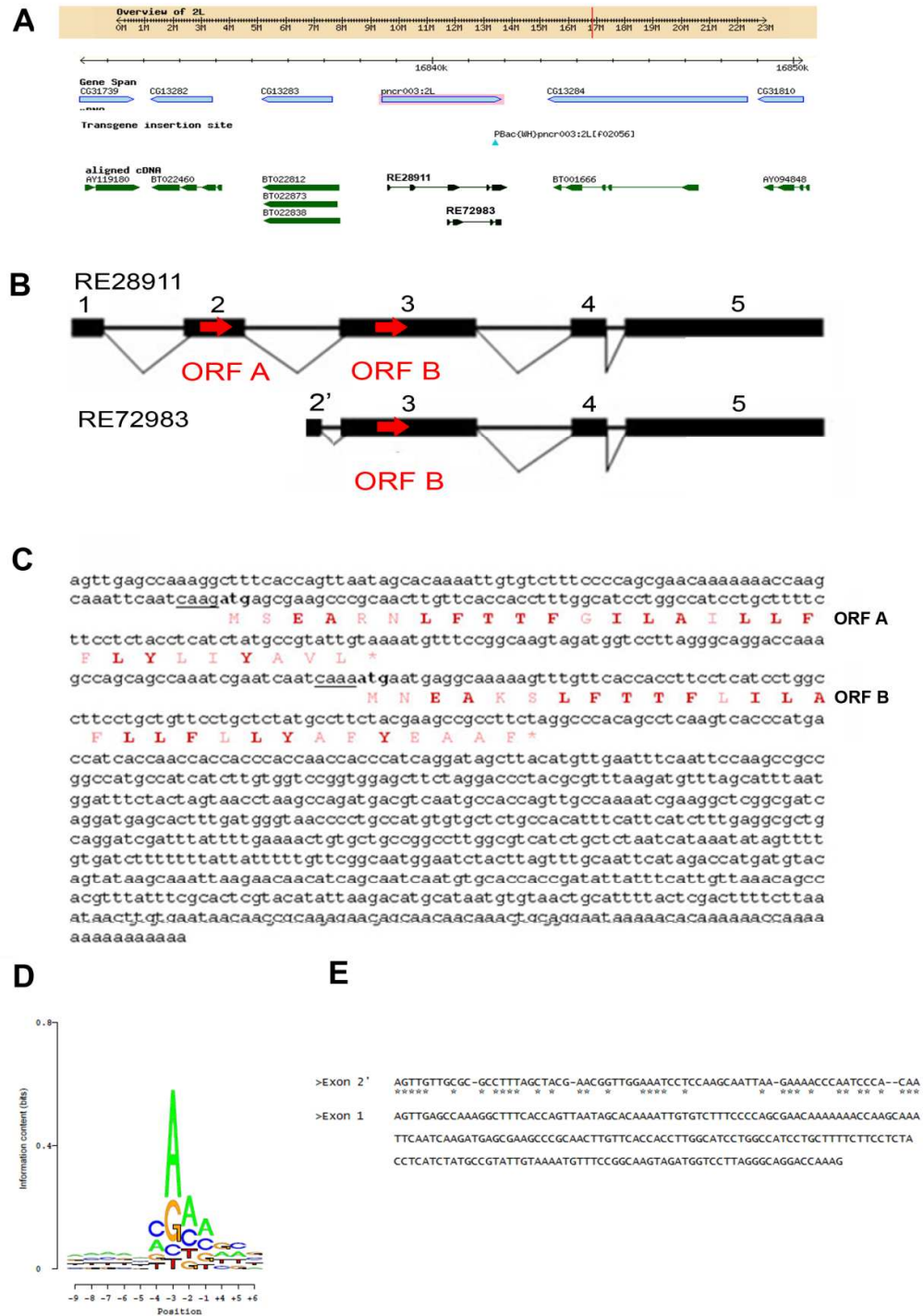


Figure 3.1

2.2- Expression of the *pncr003;2L* transcripts.

In order to assess the temporal expression of *pncr003;2L*, I carried out reverse semi-quantitative RT-PCRs on total RNA extracts from whole individuals at different stages of the *Drosophila* life cycle (Figure 3.2A). The stages assessed were: embryos, the three larval stages (L1, L2 and L3), a mixture of all pupal stages, and adults. A set of forward and reverse PCR primers (primers AB, A) were designed to anneal to exons 2 and 5 respectively, in order to amplify an 800 base pairs (bp) fragment corresponding to the transcript associated with the RE98911 cDNA. A second set of primers (primers B) was designed to anneal to exons 2' and 5, in order to confirm the existence of the alternative transcript associated with the RE72983 EST (Figure 3.2B).

The profile obtained using the first set of primers (primers AB, A) was unexpected, since all stages showed three bands of 800, 700 and 500 bp (Figure 3.2A). After sequencing, these fragments were determined to correspond to three different mRNA isoforms of *pncr002;2L* (Figure 3.2B): The 800 bp band corresponds, as expected, to the same isoform as RE28911 in which both exons 2 and 3, and thus ORFs A and B, are present; this transcript was named AB. The 700 bp band corresponds to another isoform, which also includes both exons 2 and 3, but excludes exon 4; this transcript was named AB'. The 500 bp band corresponds to an isoform which excludes exon 3, and therefore only encodes ORF A; this transcript was named A. The second set of primers, (primers B) gave rise to a single band of the expected size of 800 bp, which corresponds to the isoform associated with the RE72983 EST. Because this isoform only contains ORFB, it will be referred to as the B isoform. Although all the transcripts are present in all developmental stages, the AB and AB' transcripts show a stronger expression in embryonic stages, which diminishes in the subsequent larval and pupal stages, and appears quite weakly expressed in adults. The B transcript is more consistently expressed across the different developmental stages. The second set of primers produce the expected 800 bp band, which confirms the existence of the alternative isoform which uses exon 2'. This band is also present during all the developmental stages tested, but is also quite weakly expressed in adults.

The differential expression of these alternative transcripts, which have different constitutions with respect to ORFs A and B, is interesting as it could provide a means to modulate the expression of these two smORFs across the life span of the fly. All of these isoforms are also consistent with a local duplication of the two first exons of the

original gene, which would be represented by the isoform A. The duplicated exon 3 can be integrated into that original transcript to generate the AB transcript. Likewise the B isoform replicates the original transcript, but uses the duplicated exons instead (Figure 3.2C).

Having determined that overall these *pncr003;2L* transcripts and their respective ORFs A and B are expressed during all stages of the fly, their spatial patterns of expression of *pncr003;2L* were assessed by *in situ* hybridisation performed in whole embryos, and flat preparations of stage three larvae and adult abdomens, using a probe which anneals to exons 4-5, and therefore to all isoforms of *pncr003;2L*.

In embryos, strong expression of *pncr003;2L* can be observed exclusively in the somatic muscles, which confirms the observations of Tupy *et al.* [24] (Figure 3.2D, F). Interestingly, this expression only manifests at the latest stages of embryogenesis (stages 13 onwards) at which point the muscles have already been determined, and fully differentiated [38] (Figure 3.2E). This relatively late onset of expression in muscles already suggests that these smORFs may have a function in mature muscles, rather than during their development.

In the larva, the expression of *pncr003;2L* can still be detected in all somatic muscles, but at this stage, it is also present in cardiac muscles (Figure 3.2G). Interestingly, in adult abdominal regions the expression of *pncr003;2L* appears to be confined to cardiac muscles, as no expression could be detected in the abdominal somatic muscles (Figure 3.2H).

Figure 3.2: Transcriptional expression of *pncr003;2L*.

(A) Transcript expression profile obtained by RT-PCR at different stages of the life cycle of *Drosophila*. E: embryo; 1-3: first to third instar larvae; P: pupae (combination of all pupal stages); A: adult. The size of the bands, and the isoform they correspond to are indicated on the left of the gel. The primers used are indicated in purple (B) Diagram of the different isoforms obtained in the RT-PCR profile, indicating their exon (black rectangles) and ORF (red squares, labelled A or B) constitution, and the size of their expected PCR product (green lines). The position of the different sets of primers (primers AB, A or primers B) used in these PCRs are indicated by the purple arrows at the bottom of the diagram. (C) Diagram showing the conceptual duplication of exons 1 and 2, leading to exons 2' and 3. This model suggests that isoform A depicts the ancestral state of the gene (black squares and lines), and the duplication leads to the other alternative isoforms (orange squares and lines). Note that this duplication could have also taken place the other way around, with exons 2' and 3 being duplicated into exons 1 and 2. (D-H) in situ hybridisation, obtained with a probe annealing to exons 3-5 showing the expression of *pncr003;2L* in: (D) a lateral view of stage 17 embryos, showing expression in all somatic muscles (compare with the right panel showing the pattern obtained by driving GFP with the muscle driver *DMef2-Gal4*; (E) Lateral view of stage 14 embryo, showing no signal. (F) stage 17 Embryos showing signal in somatic muscles but not in the dorsal vessel (arrow). (G) Dorsal flat preparation of an L3 larva, showing expression in somatic (arrow head) and heart (arrow) muscles; (H) flat preparation the adult dorsal abdominal cuticle and associated muscles, showing expression of *pncr003;2L* in heart muscles (arrow) but not somatic muscles.

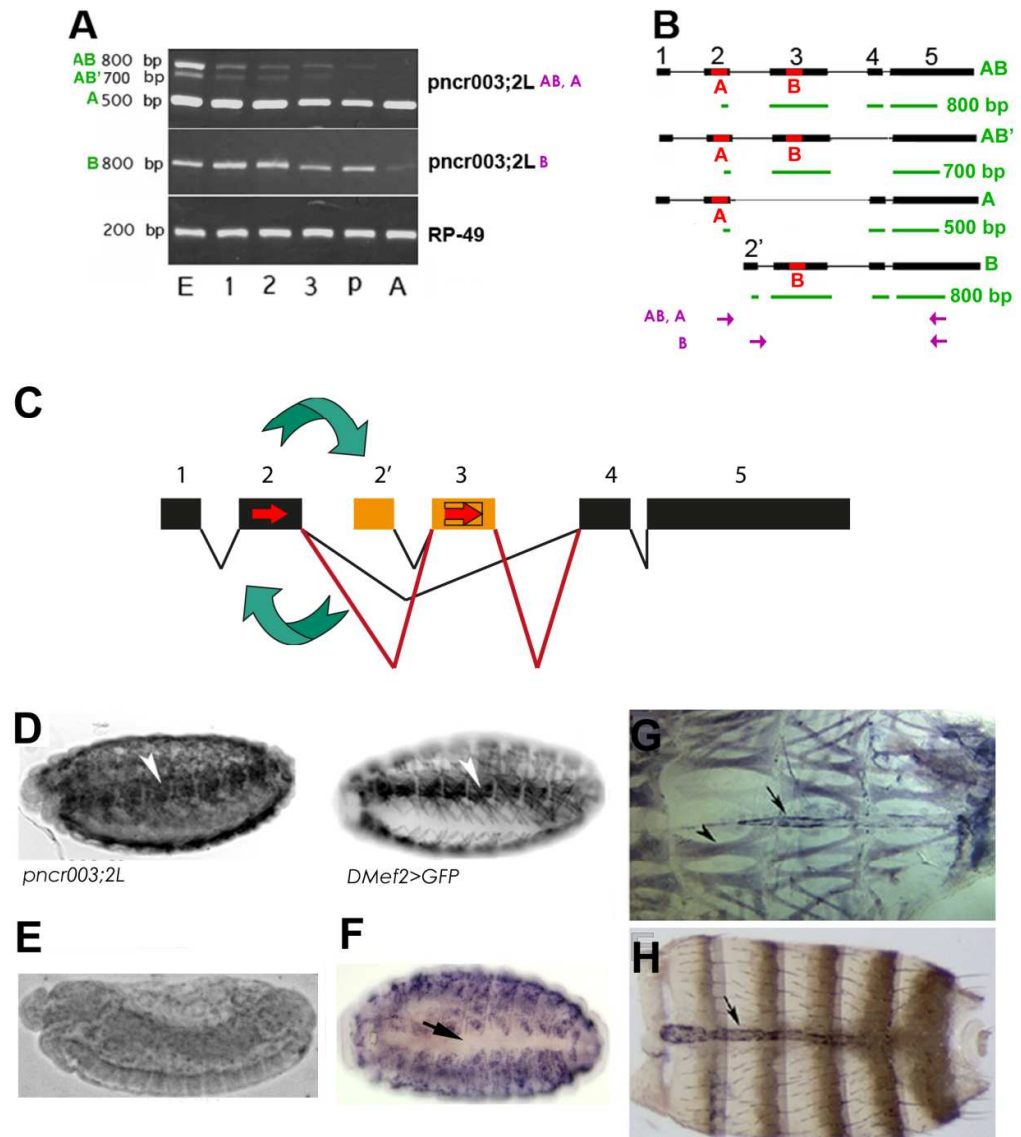


Figure 3.2

In order to determine whether *pncr003;2L* is expressed in the adult indirect flight muscles (IFMs), which follow a different developmental process to all other adult somatic muscles, I performed RT-PCRs with RNA extracted from hemi-thoraces devoid of all extremities and guts, leaving IFMs as the major contributors of RNA. The obtained RT-PCR profile shows the presence of the three AB, A and B isoforms (Figure 3.3A) indicating that *pncr003;2L* is also expressed in adult IFMs. In these muscles, all three isoforms are expressed in relatively similar proportions. The expression of *pncr003;2L* in IFMs is also supported by *in situ* hybridisation experiments (Figure 3.3B and C), which show that although high levels of background signal can be observed with the non-specific sense probe, the anti-sense probe produces a stronger signal.

As stated above, there appears to be a differential expression of the different *pncr003;2L* transcripts during the development of the fly. In order to test whether there is also tissue specific expression of these isoforms, the IFM profile was compared with the profile obtained from RT-PCRs performed with RNA extracted from adult hearts. Interestingly the heart expression profile is quite different from that of IFMs; in hearts the A isoform is completely absent, and the B isoform is expressed at much higher levels than the AB isoform, and that the B isoform itself in IFMs. These results, therefore show that there is indeed a tissue specific regulation of the expression of the different *pncr003;2L* isoforms.

Figure 3.3: The *pncr003;2L* transcripts are expressed in the Indirect Flight muscles, and show tissue specific expression. (A) Transcript expression profiles obtained by RT-PCR on extracts from adult thoraces devoid of heads, appendages and guts, leaving the indirect flight muscles (IFMs) as major contributors of tissue; or from heart extracts. The different isoforms, and primers are indicated as in Figure 3.2B. Notice the differences in the profiles between each tissue, notably, the absence of isoform A, and higher expression of isoform B in hearts. (B-C) *in-situ* hybridisation experiments showing that: (B) A strong signal in transversal (arrows) and longitudinal (arrow heads) indirect flight muscle is obtained when an anti-sense probe specific to *pncr003;2L* is used. (C) Only background signal is observed with the non-specific sense probe is used in the thoracic IFMs.

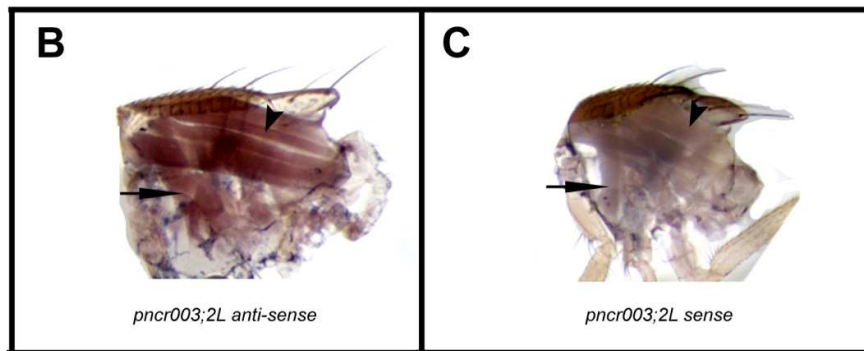
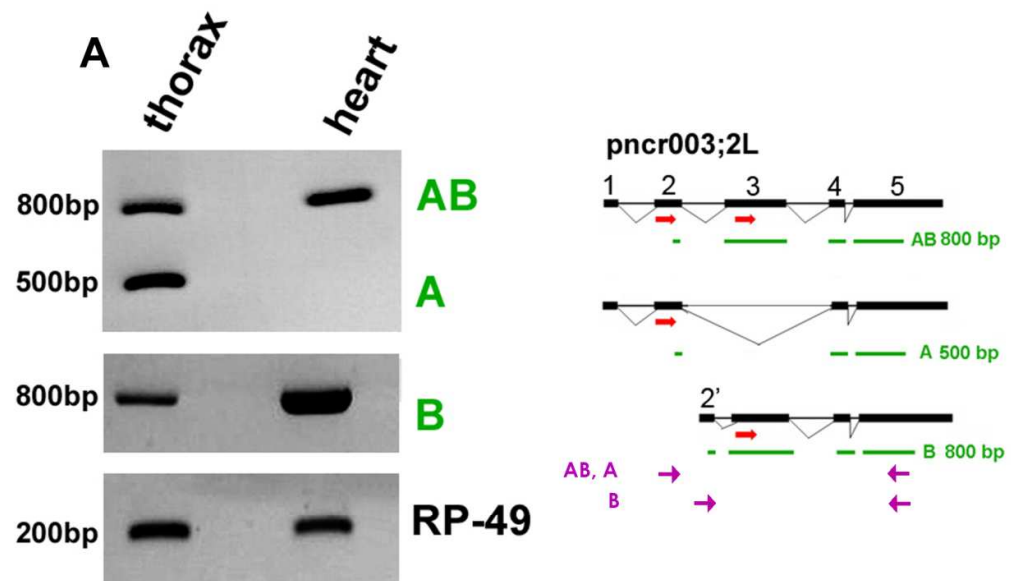


Figure 3.3

2.3- Translation of the *pncr003;2L* peptides.

To prove that the *pncr003;2L* ORFs A and B are translated, the ORFs were tagged with GFP and transfected into S2 R+ culture cells. Two constructs were made where the GFP tag, devoid of its own start codon, was cloned in frame with either ORF A or ORF B at their C-terminus (Figure 3.4). This was accomplished by introducing a unique restriction site at the C-terminus of either ORF, using the circular vector carrying RE28911 as a template for a PCR reaction with primers that annealed to the C-terminus of the ORF in opposite directions, and which replaced the stop codon of the ORFs by the unique restriction site. This PCR reaction produced a linear product, which upon ligation of the GFP sequence was recircularised, giving rise to the tagged ORF constructs. It was important to implement this kind of strategy as it led to the tagging of the ORF while preserving the whole transcript sequence, including its 5' and 3' UTRs, therefore if translation is observed one can be sure that it is because of the endogenous context of the ORFs, rather than that of an artificial expression vector. In order to allow the tissue-specific expression of these *pncr003;2L* ORF A-GFP and *pncr003;2L* ORF B-GFP constructs in transgenic flies, they were placed under the control of the widely used upstream activating sequence (UAS) promoter, which responds specifically to the yeast *Gal4* transcription factor [39]. For this, the constructs were cloned into the pUAST attB vector, which also allowed their integration into transgenic flies into a specific locus [40].

Upon co-transfection into S2 R+ cells with an *actin-Gal4* expression driver vector, a strong immunofluorescent GFP signal can be observed for both constructs (Figure 3.5A-B). Both ORFs appear to be translated at similar levels, as the intensity of their fluorescence signal is comparable. Importantly, both peptides localise to membrane structures, including the plasma membrane and perinuclear endoplasmic reticulum (ER). This membrane localisation is indicated by the strong co-localisation between ORFA and the membrane bound RFP marker (mCD8-RFP) (Figure 3.5A-A'). These results are therefore in complete agreement with the predictions presented in the introduction of this thesis, stating that these ORFs were very likely to be translated, while also showing that these peptides have a specific subcellular localisation.

Figure 3.4: Cloning strategy to assess the translation of the *pncr003;2L* ORF A and ORF B peptides. Diagram representing the ORF tagging strategy implemented to preserve the endogenous transcript sequences, and therefore the endogenous translation contexts of *pncr003;2L* ORFA and ORFB: The GFP sequence devoid of start codons was cloned in frame at the C-terminus of either ORF. For this a unique HindIII restriction site was introduced at the 3' and 5' termini of the GFP sequence by PCR, using a forward primer that removed the ATG. The HindIII site was introduced at the 3' end of either ORF A or ORF B by PCR, using a circularised vector as template with primers facing opposite directions, which also removed the stop codon of the ORFs. The *pncr003;2L* ORF A and ORF B constructs were then cloned into the pUASattB vector, which was used to transfect S2 R+ cells, or to generate transgenic flies, using site specific p-element transgenesis on the 86F8 FlyC31 *Drosophila* strain. The annealing sites of the primers are indicated by black arrows.

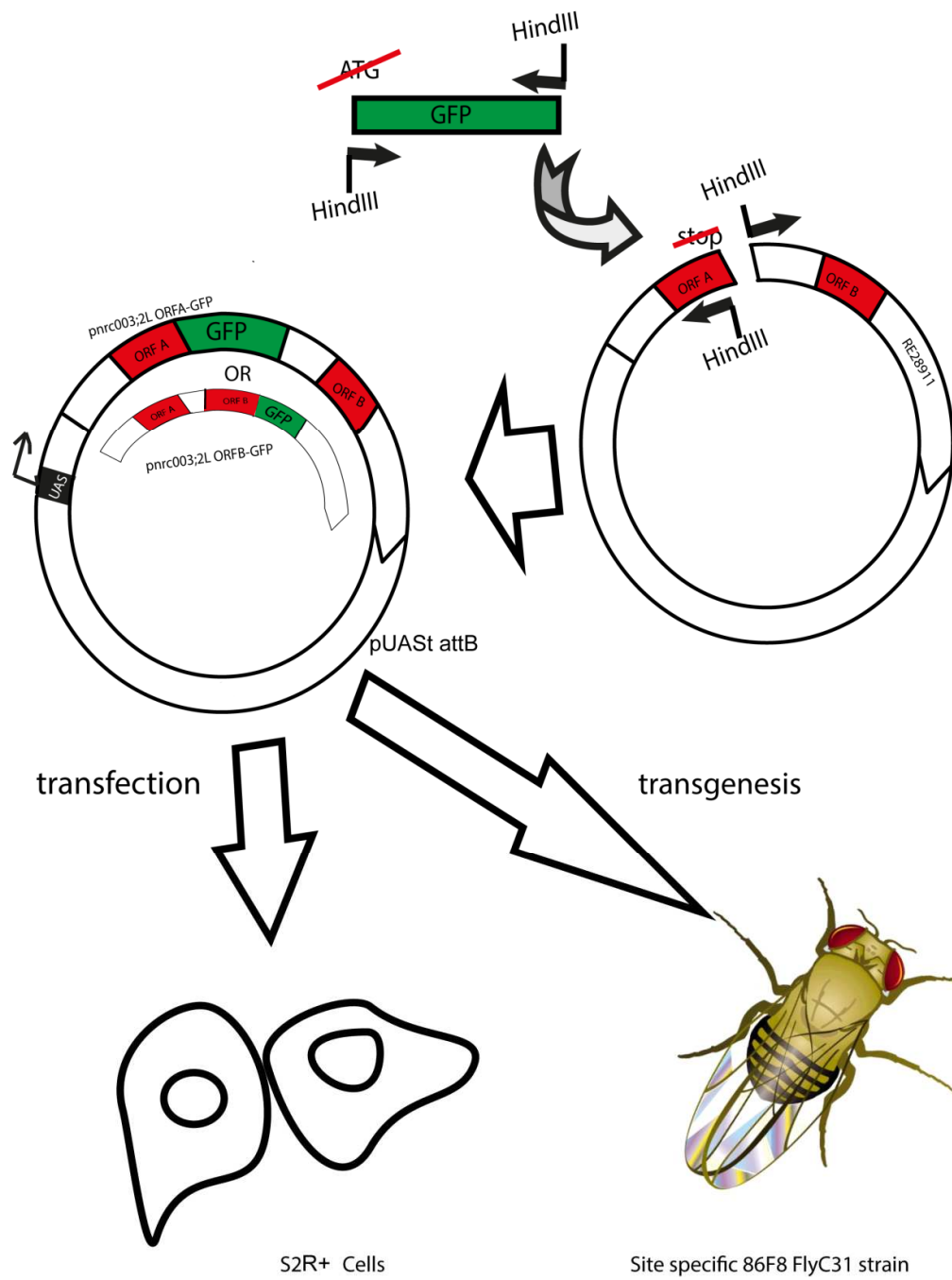


Figure 3.4

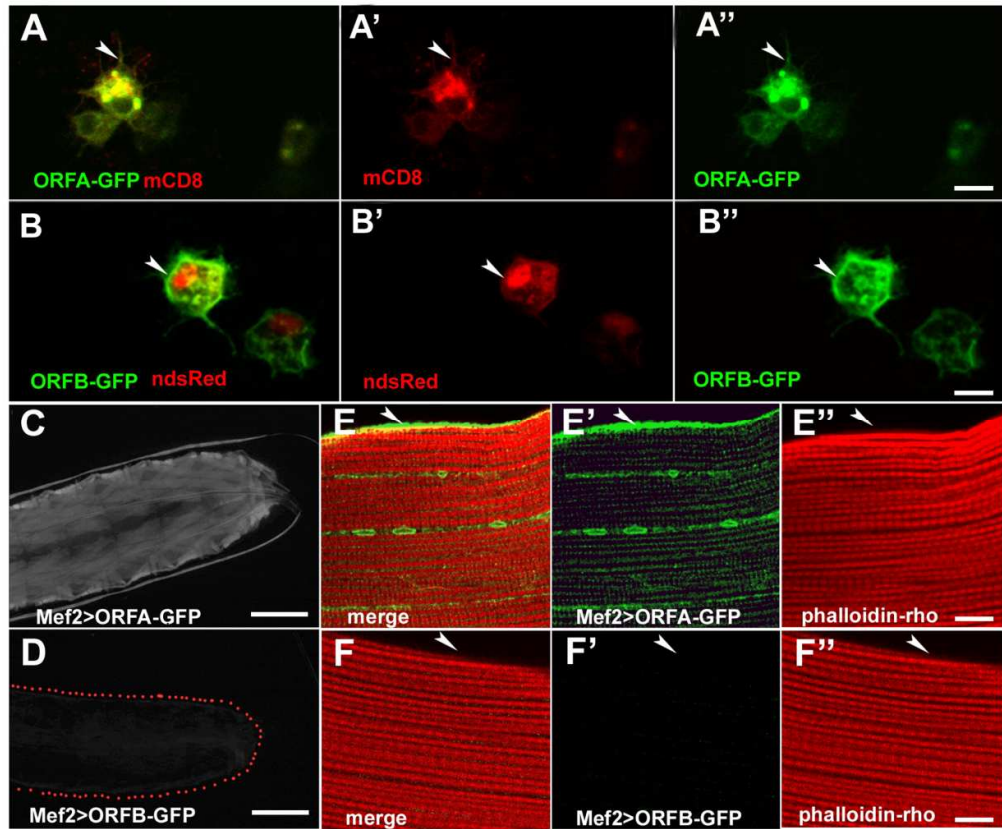


Figure 3.5

Figure 3.5: ORF A and ORF B translation in cultured cells and muscles. (A-B) S2R+ *Drosophila* culture cells transfected with (A) pncr003;2L ORF A-GFP (A'') and the membrane bound marker mCD8-RFP (A'); or (B) pncr003;2L ORF B-GFP (B'') and nuclear Ds-Red (B') arrows show localisation to the membrane. (C-D) Epi-fluorescence images of transgenic L1 larvae expressing either (C) pncr003;2L ORF A-GFP or (D) pncr003;2L ORF B-GFP. Red dots outline the larva no fluorescent signal. (E-F) Adult IFMs expressing either (E) pncr003;2L ORF A-GFP or (F) pncr003;2L ORF B-GFP. Myofibrils are labelled with phalloidin rhodamine. Arrows indicate the localisation to the plasma membrane. Scale bars: 15 μ m. All images, except (C) and (D) were acquired by confocal microscopy. Expression of both constructs in transgenic animals is driven in muscles by *DMef2-GaL4*. Scale bars: (A-B'')= 5 μ m; (C,D)= 100 μ m; (E-F'')= 15 μ m.

2.4- Subcellular localisation of the *pncr003;2L* peptides *in situ*

In order to assess their translation *in vivo*, and *in situ*, these constructs were integrated into flies through site-directed P-element transgenesis [40]. The expression of these constructs was then driven in the muscles of the transgenic flies, using *GaL4* under the *DMef2* promoter. In these conditions, the *pncr003;2L* ORF A-GFP construct produces a strong GFP signal in muscles, readily detected from the embryonic stage (not shown) and throughout the life cycle of the fly (see below). Furthermore the signal also appears to localise to the plasma membrane and perinuclear, ER-like structures (Figure 3.5E). The *pncr003;2L* ORF B-GFP construct, however, does not produce enough signal to be detected beyond background levels by fluorescence microscopy. A comparison between the GFP signal in either whole larvae, or adult indirect flight muscles, show the stark difference of expression between the two ORFs in either stage (Figure 3.5C - F). These results indicate that the polycistronic translation observed in S2 R+ cells does not take place in muscles, which suggests that some differences in the mechanisms that regulate this atypical kind of translation exist between S2 R+ cell cultures and muscles. Furthermore this inability to translate ORF B within a polycistronic transcript may also justify, or stem from, the existence of the alternative isoform B, which lacks the upstream ORF A, therefore allowing the translation of ORF B. In agreement with this, transgenic animals expressing an ORF B-mCherry construct obtained with the same tagging procedure as described above, but using the B isoform instead of the AB isoform (leading to the *pncr003;2LB* ORF B-mCherry construct), show similar levels of translation to those observed by the *pncr003;2L* ORF A-GFP construct (Figure 3.6A). The ORF A and ORF B peptides also show a great extent of co-localisation (Figure 3.6B).

Within the muscle fibres, the *pncr003;2L* peptides seem to localise in a specific pattern, of small punctate structures that outline the myofibrils. This pattern is apparent in larval somatic muscles (Figure 3.6A). In adults, a very similar pattern can be observed in the heart (Figure 3.7A), but it is more obvious in adult IFMs, which have much larger myofibrils, and therefore, offer a better resolution of these structures (Figure 3.7B).

In IFMs, the pattern appears to be quite consistently that of approximately four punctate structures per sarcomere, located between the M and Z lines. A very similar pattern has been previously described in *Drosophila* adult IFMs for the Ryanodine receptor (RyR)

[41] (see inset in Figure 5.7C). The RyR is the intracellular calcium channel responsible for the release of calcium from the sarco-endoplasmic reticulum (SER) into the cytosol, which triggers muscle contraction. In cardiac muscles these channels are activated by calcium itself in a process known as calcium-induced calcium-release (CICR). The local amounts of Calcium ions necessary for CICR, are provided from the extracellular space through L-type voltage mediated calcium channels (L-CaCh) located at close proximity from the RyRs. In skeletal muscles, the L-CaCh are in physical contact with the RyR, and activate them directly. The RyRs are therefore located in a region of the SER which is in close apposition to an invagination of the muscle cell plasma membrane, which extends deeply into the muscle cell, allowing the transmission of the depolarising currents necessary for the activation of the L-CaCh all the way from the neuromuscular junction to the myofibril-associated SER (Figure 3.8). The junction between these membrane invaginations, known as T-tubules and the SER, is known as dyad (Figure 3.7C, Figure 3.8), a structure which is at the core of the muscle excitation-contraction coupling system.

Although, unfortunately no anti-RyR antibody could be obtained to confirm the co-localisation between the RyR and the pncr003;2L peptides, it could be determined that these peptides largely co-localise with the T-tubule system, detected with anti *Discs-large*, as described in [41] (Figure 3.7D).

This very specific subcellular localisation, which points to a physiological role for these peptides, is not an artefact of the GFP tagging method used, since it can be reproduced with N-terminal FLAG-Hemagglutinin tagged peptides (FH-ORF A and FH-ORF-B) — generated by Dr. Jose Pueyo [33]—, a completely different tag because of its nature, smaller size, and N –terminal localisation with respect to the ORF A and ORF B peptides (Figure 3.9A and B). Furthermore, this precise subcellular localisation is not due either, to their membrane bound structure. Indeed, other membrane bound markers, such as mCD8-RFP fail to reproduce the pattern observed with the pncr003;2L peptides (Figure 3.9C). This evidence suggests that the tagged peptides reflect the localisation of the endogenous peptides.

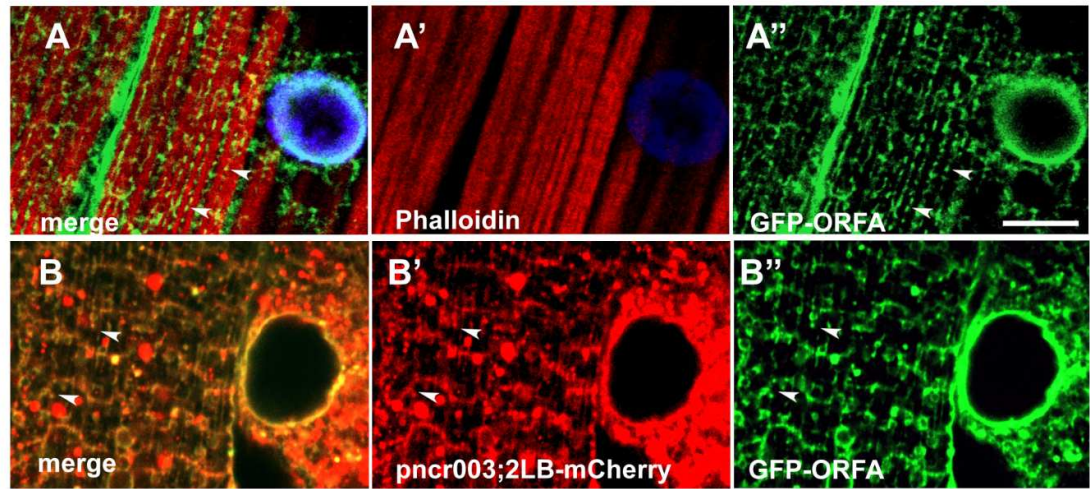


Figure 3.6

Figure 3.6: ORF B is translated from the B isoform in situ, and has a similar expression and localisation as ORF A. (A-B) Larval somatic muscles expressing (A) pncr003;2L ORF A-GFP (A'') and stained with phalloidin rhodamine (A'); or (B) pncr003;2LB ORF B-mCherry (B') and pncr003;2L ORF A-GFP (B''). Arrows indicate the localisation to the dotted structures outlining the myofibrils. A DAPI stained nucleus (blue) is shown in (A). The expression of these constructs is driven in muscles by *DMef2-GaL4*. All images were acquired by confocal microscopy. Scale bar: (A-B'')= 15µm.

Figure 3.7: The pncr003;2L peptides localise to the dyads. (A-B) expression of pncr003;2L ORF A-GFP (green) in (A) adult heart and (B) IFMs. All samples are counterstained with phalloidin-rhodamine (red or cyan) to label the myofibrils. DAPI stained nuclei are shown in blue. (C) Transmission electron micrograph of an IFM sarcomere showing the localisation of the dyads. The inset, (modified from Razzaq *et al.* [41]) shows the localisation of the dyad associated Ryanodine receptor on a magnified confocal fluorescence image of an IFM sarcomere. (D) Expression of pncr003;2L ORF A-GFP in IFMs counterstained with T-tubule marker *Discs-large* (*Dlg*). Arrow heads point to dyads. Scale bars: (A, B, D)= 10 μm ; (C)= 0.4 μm .

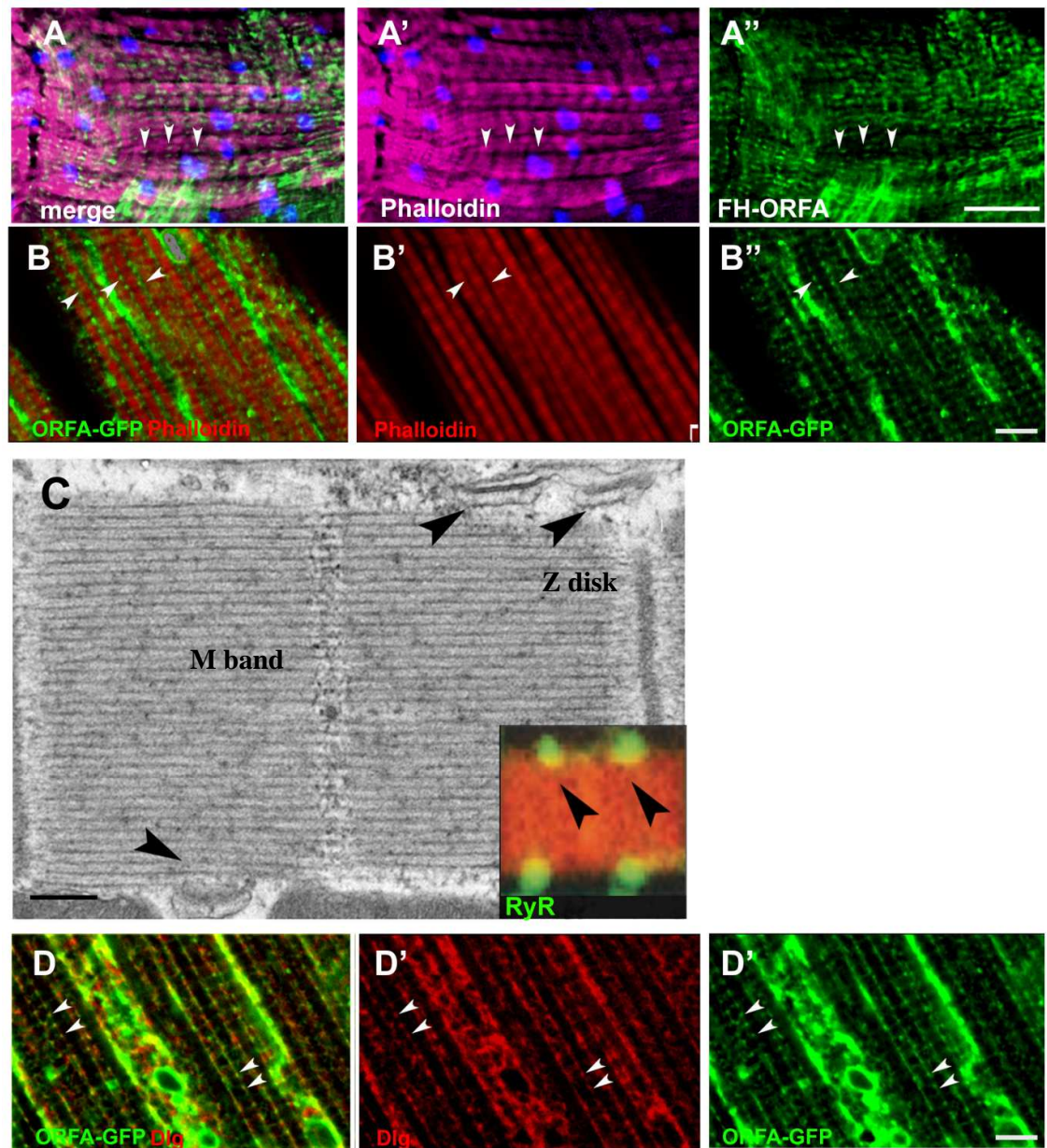


Figure 3.7

Figure 3.8: The dyad is at the core of calcium regulation and muscle contraction.

Diagram of calcium trafficking during muscle contraction modified from Gwathmey J.K *et al.*, 2011 [42]. Dyads are the cellular structures where the plasma membrane (T-tubules) is in close apposition to the sarcoendoplasmic reticulum membrane (faded red ellipse). Muscle contraction is regulated by the intracellular levels of calcium. When the neuronal membrane potential is transmitted into the muscle cell, the L-type voltage activated calcium channels transport extracellular calcium into the cytoplasm. This calcium increase triggers the release of large amounts of calcium from the sarcoplasmic reticulum (SER) through the Ryanodine receptor (RyR) channels, inducing the contraction of the sarcomeres. Muscle relaxation is achieved by the depletion of calcium from the cytoplasm, through the sarcoendoplasmic reticulum calcium ATPase (SERCA), while sodium/calcium exchange pumps (NCX) and plasma membrane calcium ATPase pumps (PMCA) release calcium out into the extracellular space (the percentages indicate the estimated contribution of each pump towards the depletion of cytoplasmic calcium).

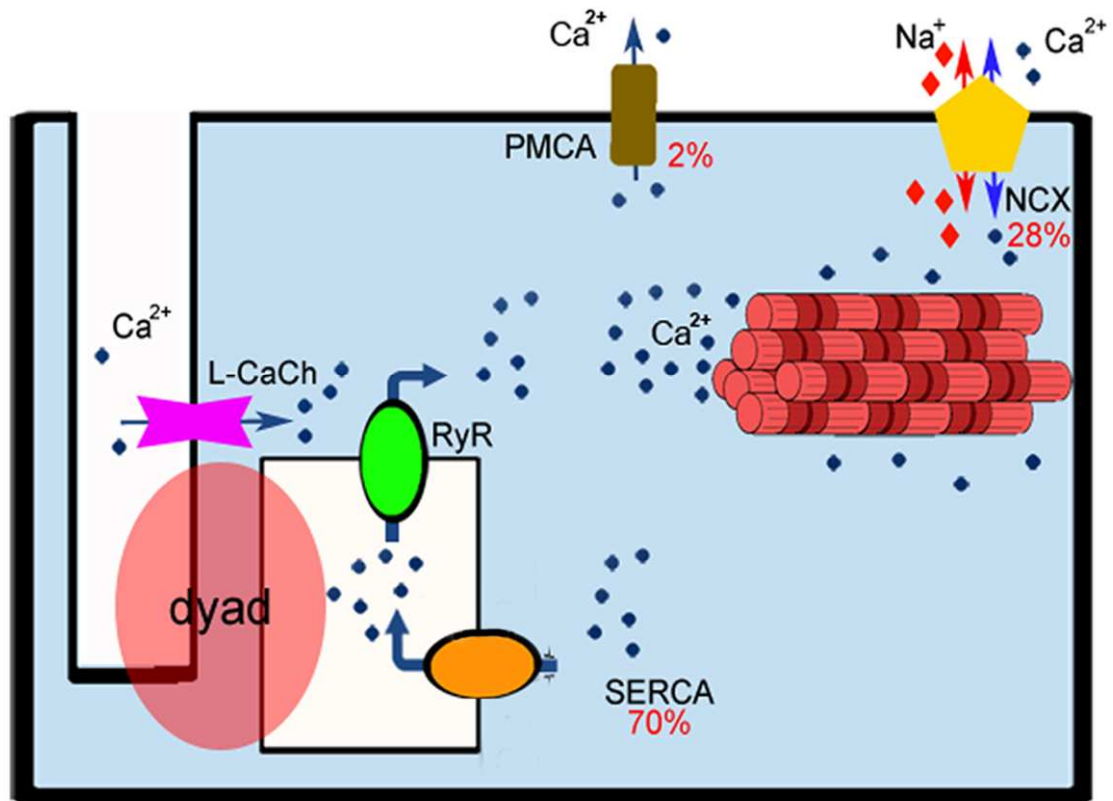


Figure 3.8

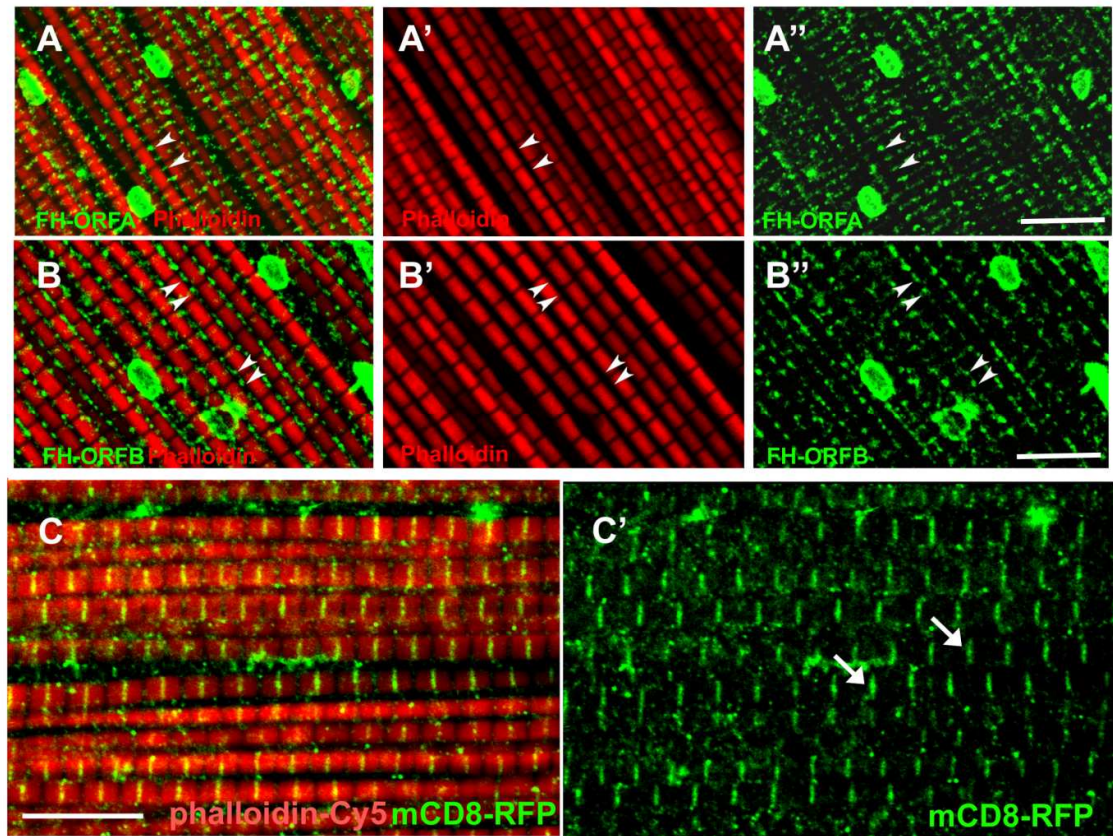


Figure 3.9

Figure 3.9: The subcellular localisation of *pncr003;2L* ORF A and ORF B is not artifactual. Expression of (A) *pncr003;2L* FH-ORF A and (B) *pncr003;2L* FH-ORF B, kindly provided by J.I. Pueyo-Marques [33], in adult IFMs counterstained with phalloidin rhodamine. Both peptides show the same localisation to the dyads as described in (Figure 3.8) (arrows). (C) Membrane bound RFP (mCD8-RFP, in green) shows a different localisation pattern with no particular enrichment in dyads, the signal seems to localise to z discs instead (arrows). Scale bars (A-C')= 10µm.

3- Discussion:

This work provides the basis for the functional characterisation of *pncr003;2L*, while portraying a completely different picture of *pncr003;2L* compared with what has been previously described, particularly because it is shown here that *pncr003;2L* is a coding gene, rather than a non-coding RNA, because it codes for two small ORFs of 28 and 29 codons that are translated into peptides, and which can therefore be considered as *bona fide* smORFs. These smORFs are expressed in muscle tissues throughout the life cycle of *Drosophila*.

The identification of a second coding ORF in the *pncr003;2L* gene, highlights another obvious unusual feature of *pncr003;2L* and *tarsal-less*, apart from the short size of their encoded peptides, which is their polycistronic nature. Although polycistronic genes are common in prokaryotes, they are very rare in most eukaryotes, with the *Trypanosome* and nematode genomes, being the only examples of eukaryotes where polycistronic transcripts have been widely detected. In *Caenorhabditis elegans* for example, at least 2600 genes, representing 15% of the total number of genes in that organism, have been determined to belong to operons [43]. In other eukaryotes, only a handful of polycistronic transcripts have been described. In *Drosophila* for example, the gene coding for the Stoned A and B peptides (*stn A and stn B*), as well as the gene coding for the alcohol dehydrogenase and alcohol dehydrogenase related proteins (*Adh and Adhr*), are both known to produce di-cystronic transcripts encoding these polypeptides [44]. Phylogenetic analysis of *stn* and *adh* genes show that the polycistronic arrangement is the primitive state of the gene [44], whereas the high amino acid conservation between ORF A and ORF B, and between the *tal* smORFs, suggests that in the case of *pncr003;2L* (and also in the case of *tal*) this polycistronic arrangement may have been acquired upon a local duplication of the original locus.

The apparent inability of dicistronic translation, *in situ*, for this gene in the tissues studied, together with the different alternative transcripts detected here, which in some cases, encode for only one of two smORFs, and have independent promoters, leading to differential tissue specific expression for each transcript, and therefore for each smORF, suggests that the case of the *pncr003;2L* ORFs A and B resemble more that of two separate paralogous genes, sharing the terminal region of their transcript by the

proximity of their duplication. In the case of *tal*, a similar duplication also seems to be at the origin of the observed polycistronic transcript. However, in that case, the lack of splicing in the original transcript may have forced the translation of the polycistronic ORFs. Because of the extensive sequence similarity of the *pncr003;2L* peptides on the one hand and *tal* peptides on the other, it is unlikely that the respective members of each of these two families of peptides have functions that differ dramatically from each other. This has in fact been shown for *tal*, where each of the individual peptides is able to reconstitute the function of the whole polycistronic gene [11]. In the case of *pncr003;2L* however, the observed tissue-dependent differential expression of each ORF may allow these sequences to be in some sort of adaptive evolutionary state, in which case subtle differences in their function may eventually, if not already, be observed.

From their late onset of expression and their prevalence in muscles it could be inferred that their role would take place in mature muscles rather than in the developmental process of muscles. This functional prediction is reinforced by the very specific subcellular localisation of the peptides to the dyads, which suggests that these peptides may have some sort of role, like other proteins localised to those structures, in the physiological processes regulating muscle contraction. The fact that this subcellular localisation is independent from their ability to localise to the plasma membrane in S2R+ cells and muscle tissues, because other plasma membrane bound markers such as mCD8-RFP do not exhibit it, indicates that something else, possibly a functional partner, must be stabilizing the peptides to that position.

Chapter IV - The *pBac {WH}F02056* insertion, and its use to generate null mutants for *pncr003;2L*

1- Introduction:

To ascertain the *pncr003;2L* gene function in muscles and carry out a phenotypical characterisation of this gene, I needed to generate null mutants. Although, as a first approach, one could have considered using RNA interference [45], which is arguably easier to implement than conventional mutagenesis protocols. It is important to keep in mind that RNAi is considered a “knockdown” rather than “knockout” approach [46]. Therefore, residual activity of the targeted gene may prevail, and this may be sufficient to conceal its mutant phenotype. Moreover, RNAi can sometimes produce artifactual phenotypes unrelated to the targeted gene itself [47,48]. Consequently, any phenotype, or lack of it, obtained by inducing RNA interference against *pncr003;2L*, would ultimately need to be validated against that of a null mutant, or at least against some sort of sophisticated control, like a genetic rescue refractory to RNAi.

To generate a null mutant for *pncr003;2L*, it is possible to take advantage of a particular *Drosophila* strain which carries a transposable element (*pBac {WH}F02056*) mapped to the 3'UTR of *pncr003;2L* (Figure 4.1, Annex 3). Transposable elements are capable of mobilising within the genome, sometimes inserting themselves in regions that may disrupt the expression of their neighbouring genes. They exist in the endogenous genomic sequences of all species, and their relevance is such that it has been suggested

that they have played a role in the overall shaping of higher eukaryote genomes [49]. Their ability to transpose, or mobilise, has been widely studied in *Drosophila melanogaster*, ultimately leading to the development of genetically engineered transposons, used as highly effective mutagenic reagents [50,51,52]. Such transposons have been designed to lack the sequence coding for the Transposase enzyme, which catalyses their mobilisation, and are therefore stable in the genome, unless the Transposase enzyme also becomes available. The extensive usage of transgenic transposons has led to the creation of several randomly generated libraries of strains carrying single transposon insertions, each affecting a specific gene or genomic region [53]. Depending on where these transposons are inserted, around or within their target genes (intron, UTRs, regulatory region, or ORF), they can have a range of effects on the expression of the gene, ranging from no effect at all (usually associated with intronic or UTR insertions) to total disruption, which would give rise to a null allele for that gene. The most severe effects are usually associated with insertions in the ORF, or regulatory regions of the gene.

The *pBac{WH}F02056* insertion, comes from the Exelixis collection [54]. The transposons used to generate this particular collection were engineered to carry the FLP recombination target (FRT) sequence [55], which allows them to be used to generate custom genomic deficiencies [56]. Like most other transposons used to generate these libraries, they also carry an insertion marker, namely a constitutively expressed version of the *white* gene coding for the protein responsible for the red pigment of the eye.

In the first part of this chapter, I address the effects of the *pBac{WH}F02056* insertion itself on *pncr003;2L*. I show that this insertion, which is homozygous viable, and which is located the 3'UTR of this gene, has an effect on the levels of *pncr003;2L* transcript detected in adult flies, hence representing a hypomorphic allele for this gene. I describe an interesting phenotype detected in this strain, which manifests in the adult indirect flight muscles (IFMs), a tissue where *pncr003;2L* is expressed (Chapter III). The *pBac{WH}F02056* IFMs have myofibrils with shorter sarcomeres. Interestingly, this phenotype appears to be additive to that generated by a haploinsufficiency of the *Myosin heavy chain (Mhc)* gene, which also leads to short sarcomeres. Although this initial observation seems to suggest that a promising genetic interaction exists between *pncr003;2L* and *Mhc*, the phenotypical analysis of different combinations of *pncr003;2L* and *Mhc* genetic conditions, led me to propose that this phenotype is due to

an associated mutation carried in the second chromosome of the *pBac{WH}F02056* strain, probably affecting the *Mhc* locus itself, rather than to the effect of the insertion on *pncr003;2L*. This hypothesis was confirmed by mapping the allele responsible for the short sarcomere phenotype to the *Mhc* locus, using a combination of genetic and molecular methods.

In the second part of this chapter, I describe two mutagenesis approaches, which make use both of the *pBac{WH}F02056*. The first approach, follows an FRT recombination protocol as described by Parks *et al.* [56], slightly modified in order to facilitate the screening process for the flies carrying the putative deficiencies, to generate a custom deficiency covering *pncr003;2L* and two other genes. The second approach consists of a γ -ray mutagenesis protocol, in which I screened for γ -ray induced DNA lesions interfering with the expression of the *white* gene carried by the *pBac{WH}F02056* insertion, and therefore affecting the genomic locus of *pncr003;2L*. These methods led to the generation of two small genomic deficiencies (*Df(2L)12*, and *Df γ -ray 6*), which, as trans-heterozygous, constitute a null condition of *pncr003;2L* (*Df pncr003;2L*). These *pncr003;2L* null flies are homozygous viable as adults, and are able to move and fly normally, and consistent with this, their IFM myofibrils have sarcomeres with a normal appearance. Although initially disappointing, this lack of a morphological phenotype is consistent with the assumption that the *pncr003;2L* peptides may play a physiological, rather than structural or morphological role.

2- Results:

2.1- The *pBac {WH} F02025* line is hypomorphic for *pncr003;2L* and has a specific muscle phenotype.

As already discussed, the effect of a transposable element insertion on the function of its “host” gene can vary from none to total disruption. It was therefore important to determine whether the *pBac{WH}F02056* insertion (from now on referred to simply as *F02056*) had an effect on the expression of the *pncr003;2L* transcripts. For this, I carried out a semi-quantitative RT-PCR on whole fly RNA extracts, using PCR primers that amplify fragments corresponding to either of the two peptide coding exons 2 and 3. The bands corresponding to both exons are visibly weaker for the *F02056* homozygous

flies, than for the wild-type (wt) controls (Figure 4.2C), indicating that the *F02056* insertion does affect the expression of *pncr003;2L*, leading to a reduction of the expression level of its transcripts, and thus it behaves like a hypomorphic allele of this gene.

To look at the effects of this insertion on the motility of flies, I used a simple test to assess their flight capabilities consisting essentially of releasing individual flies on a flat surface, using a suction tube, and assessing whether they can fly away (see materials and methods, section 2.14). This assay revealed that the flies homozygous for *F02056* seem to have a subtle flight defect, since almost a third of the transposon-carrying flies tested were unable to fly, in contrast with wild type flies, which were all flight able (Table 4.1). These results led me to focus on the morphology of IFMs, which are the main muscles responsible for flight in Dipterans, and importantly, where *pncr003;2L* is expressed (Chapter III, Figure 3.3). Confocal microscopy on thoracic hemi-segments stained with phalloidin, shows visible defects in the organisation of the myofibrils in flies homozygous for the *F02056* insertion (Figure 4.2A and B). The *F02056* myofibrils have a slightly “wavy” appearance, compared to the very straight wild-type myofibrils, and often present some extent of splitting. The sarcomeres within the myofibrils also seem abnormally shorter in *F02056* homozygous flies, displaying a square-like shape, as opposed to the rectangular shape observed in wild-type myofibrils. A quantification of the length of their sarcomeres indicates that *F02056* homozygous flies have sarcomeres measuring 2.8µm in average, which is significantly shorter than wild-type sarcomeres, where the average length is 3.4µm (Figure 4.2A, B and D).

2.2- The phenotype associated with the pBac {WH} F02025 line is enhanced by a *Myosin heavy chain* haploinsufficiency.

If the observed phenotypes are indeed caused by the hypomorphic condition of *pncr002;2L*, one would expect that such phenotypes would become more pronounced if the levels of *pncr003;2L* were diminished further, by placing the chromosome carrying the *F02056* insertion over a null allele, or a chromosome carrying a deficiency for that same chromosomal region. This genetic complementation test was carried out using the deficiencies *Excel 8036* (*Df 8036*), and *ED 1153* (*Df ED 1153*) (Figure 4.3A). Although both deficiencies entirely cover the *pncr003;2L* locus they both have a completely opposite effect over the *F02056* insertion: The myofibrils from *F02056 / Df 8036* flies have a wild-type-like appearance and normal sarcomeres lengths (Figure 4.3 B and E,

compare to figure 4.2A and B), whereas the myofibrils from *F02056 / Df ED1153* flies seem to have an enhanced disrupted phenotype, displaying in particular much shorter sarcomeres than the *F02056* homozygous flies; with the *F02056 / Df ED1153* sarcomeres measuring an average of 2µm (Figure 4.3C and E). An important difference between these deficiencies is that *Df ED1153* covers the *Myosin heavy chain* (*Mhc*) locus, whereas *Df 8036* does not (Figure 4.3A). *Mhc* is one of the main components of the thick filaments within the myofibrils, and therefore a major component of the muscle contraction machinery, so much so, that *Mhc* null alleles have a dominant flightless phenotype [57] (Table 4.1, Figure 4.15D). Consistent with this, heterozygous flies for *Df Ed1153*, which essentially represents a null allele for *Mhc*, are flightless, and interestingly, have a myofibril phenotype not too dissimilar from that of homozygous flies for *F02056*, with sarcomeres significantly shorter than wild type; measuring in average 3µm (Figure 4.3D and E). Two important conclusions from these genetic experiments are: a) that because the *F02056* associated phenotype is completely corrected over a null allele for *pncr003;2L* (represented by the *Df 8036*), the *F02056* homozygous phenotype cannot be due to a hypomorphic condition of *pncr003;2L*; and b) that the enhancement of the *F02056* associated phenotype seems to require a haploinsufficient condition for *Mhc*. Furthermore, the independent phenotypes of either *Mhc* heterozygous null, and homozygous *F02056* are similar, both displaying shorter sarcomeres, and appear to be additive. Similar results indicating an *F02056* dependent enhancement of the *Mhc* phenotype could be observed when examining the flight capabilities of the flies as described above. Flies carrying a *Mhc* null chromosome are flightless, but can perform very small jumps in their attempts to take-off. When such *Mhc* null chromosomes are trans-heterozygous over the *F02056* insertion, the flies are no longer able to perform these small jumps, and are left only with the ability to walk (Table 4.1).

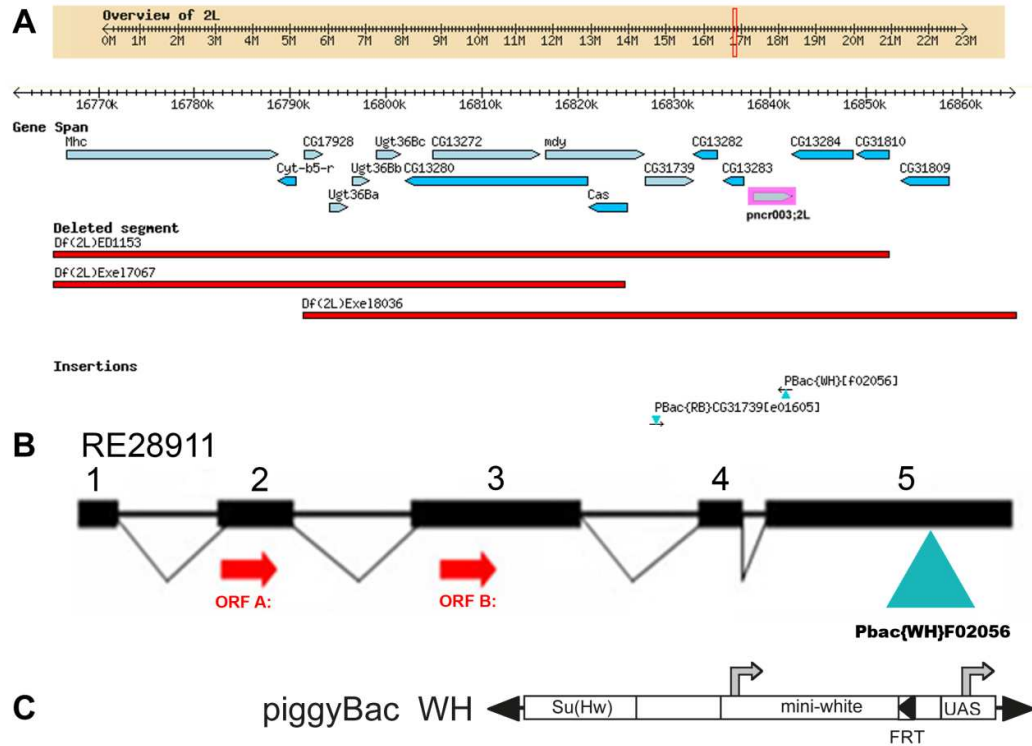


Figure 4.1

Figure 4.1: Diagram representing the genomic landscape and genetic deletions surrounding *pncr003;2L*, and the structure of the *pBac{WH}cF02056* insertion.**

(A) Diagram representing the genomic locus of *pncr003;2L* in chromosomal arm 2L, as depicted in the FlyBase genome browser. Genes are indicated by blue arrows. The *pncr003;2L* gene is highlighted in pink. Red lines represent the span of the different genomic deletions, and blue triangles the transposable element insertions, used in this study. (B) Schematic representation of the RE28911 cDNA, showing the position of ORFs A and B, and the approximate insertion site of the pBac{WH}F02056 element in the 3'UTR of *pncr003;2L*. (C) Diagram representing the different genetic elements within the pBac{WH}F02056 transposon is also represented, including the mini-white gene (*white* marker) and the FRT sequence.

| Genotype | n | percentage of flies able to : | | |
|-------------------------------------|----|-------------------------------|------|-----|
| | | Walk only | Jump | Fly |
| <i>w;Df ED1153 / +</i> | 35 | 14 | 83 | 0 |
| <i>w;Df 7067 / +</i> | 19 | 21 | 79 | 0 |
| <i>w;Df 8036 / +</i> | 20 | 0 | 0 | 100 |
| <i>w;Df ED1153 / pBac{WH}F02056</i> | 30 | 100 | 0 | 0 |
| <i>w;Df 7067 / pBac{WH}F02056</i> | 30 | 100 | 0 | 0 |
| <i>w;Df 8036 / pBac{WH}F02056</i> | 30 | 0 | 0 | 100 |
| <i>Or-R</i> | 30 | 0 | 0 | 100 |
| <i>w;pBac{WH}F02056</i> | 30 | 20 | 7 | 73 |

Table 4.1

Table 4.1: Initial characterisation of the motility of different genetic conditions affecting *pncr003;2L*. Initial characterisation of the motility of different genetic conditions affecting *pncr003;2L*, performed by a simple assay in which the flies of the indicated genotypes were released on a flat surface, and their motility scored according to three indicated categories. Flies that are haploinsufficient for *Mhc* (*w;Df ED1153 / +* and *w;Df 7067 / +*) are flightless, but able to perform small jumps in most cases, whereas flies that are haploinsufficient for *pncr003;2L* (*w;DfExel 8036 / +*) can fly normally. Over the *pBac{WH}F02056* insertion, the flightless phenotype of *Mhc* haploinsufficient flies is enhanced, as they all lose the ability to perform small jumps, but the *pncr003;2L* haploinsufficient flies can still fly normally. *pBac{WH}F02056* homozygous flies show an intermediate flightless phenotype.

Figure 4.2: The *pBac{WH}F02056* insertion gives rise to a hypomorphic condition for *pncr003;2L*, and to a short sarcomere phenotype.

(A-B) Confocal microscopy images of hemithoraces stained with phalloidin-rhodamine (red), showing the myofibrils of longitudinal indirect flight muscles of (A) *Or-R*, wild-type flies and (B) homozygous flies for the *pBac{F02056}* insertion. Nuclei are stained with DAPI (blue). The insets show a slightly magnified section of the images, representative sarcomeres are framed to highlight the differences in sarcomere shape and length between these genotypes. (C) Semi-quantitative RT-PCR on mRNA extracts of whole flies, using primers specific to either the exon 2, or exon 3 of *pncr003;2L*, showing the visible reduction in *pncr003;2L* expression in flies homozygous for the *pBac{F02056}* insertion. The primers to amplify the exon 2 and 3 fragments are represented in the diagram below the gel (purple arrows). (D) Quantification of the differences in sarcomere length between these genotypes, showing a significant difference between *Or-R* *pBac{F02056}* insertion, as indicated by a one-tailed unpaired t-test statistical analysis ($t=10.57$, $p<0.0001$) . $n=200$ sarcomeres, from 4 different flies per genotype.

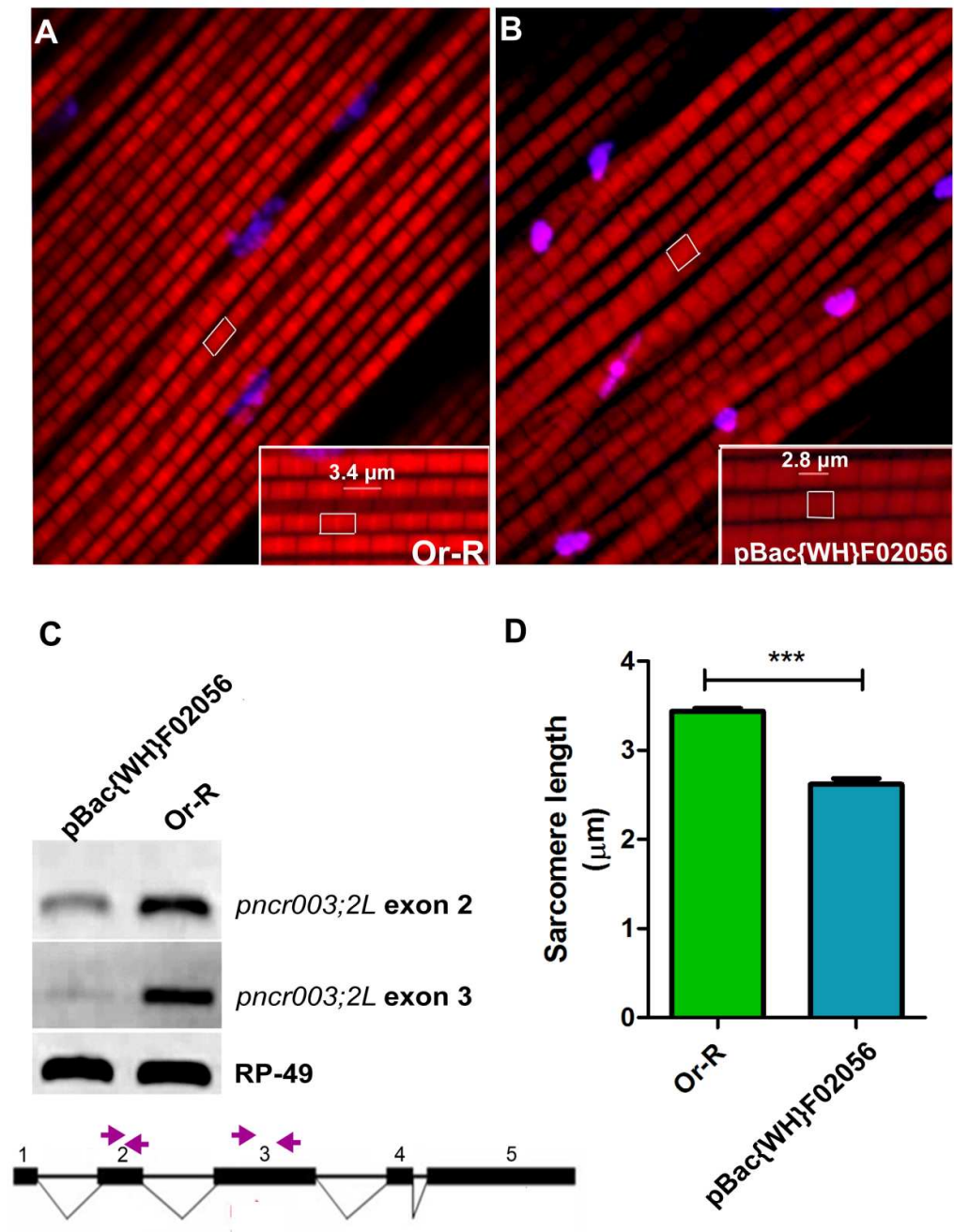


Figure 4.2

Figure 4.3: The short sarcomere phenotype associated with the *F02056* insertion

seems to be enhanced by a *Mhc* haploinsufficiency. (A) Diagram representing the genomic locus of *pncr003;2L* in chromosomal arm 2L, annotated as in Figure 4.1 A.

(B-D) Confocal microscopy images of hemithoraces stained with phalloidin-rhodamine (red), showing the myofibrils of longitudinal indirect flight muscles of (B)

Df8036/F02056 flies, with a wild type appearance (C) *ED1153 / F02056* flies, with an enhanced short sarcomere phenotype and (D) *ED1153 / +* flies with a similar short sarcomere phenotype as that of homozygous flies for the *F02056* insertion.

Representative sarcomeres are framed to highlight the differences in sarcomere shape and length between these genotypes. (E) Quantification of the differences in sarcomere length between these genotypes, showing a significant difference, as indicated by a one-tailed unpaired t-test statistical analysis, between Or-R and homozygous flies for the *F02056* insertion ($t=10.57$, $p<0.0001$), between homozygous flies for the *F02056* insertion and *ED1153 / F02056* flies ($t=5.97$, $p<0.0001$), and between Or-R and *ED1153 / +* flies ($t=4.72$, $p=0.0005$). $n=200$ sarcomeres, from 4 different flies per genotype. Scale bar= $10\mu\text{m}$.

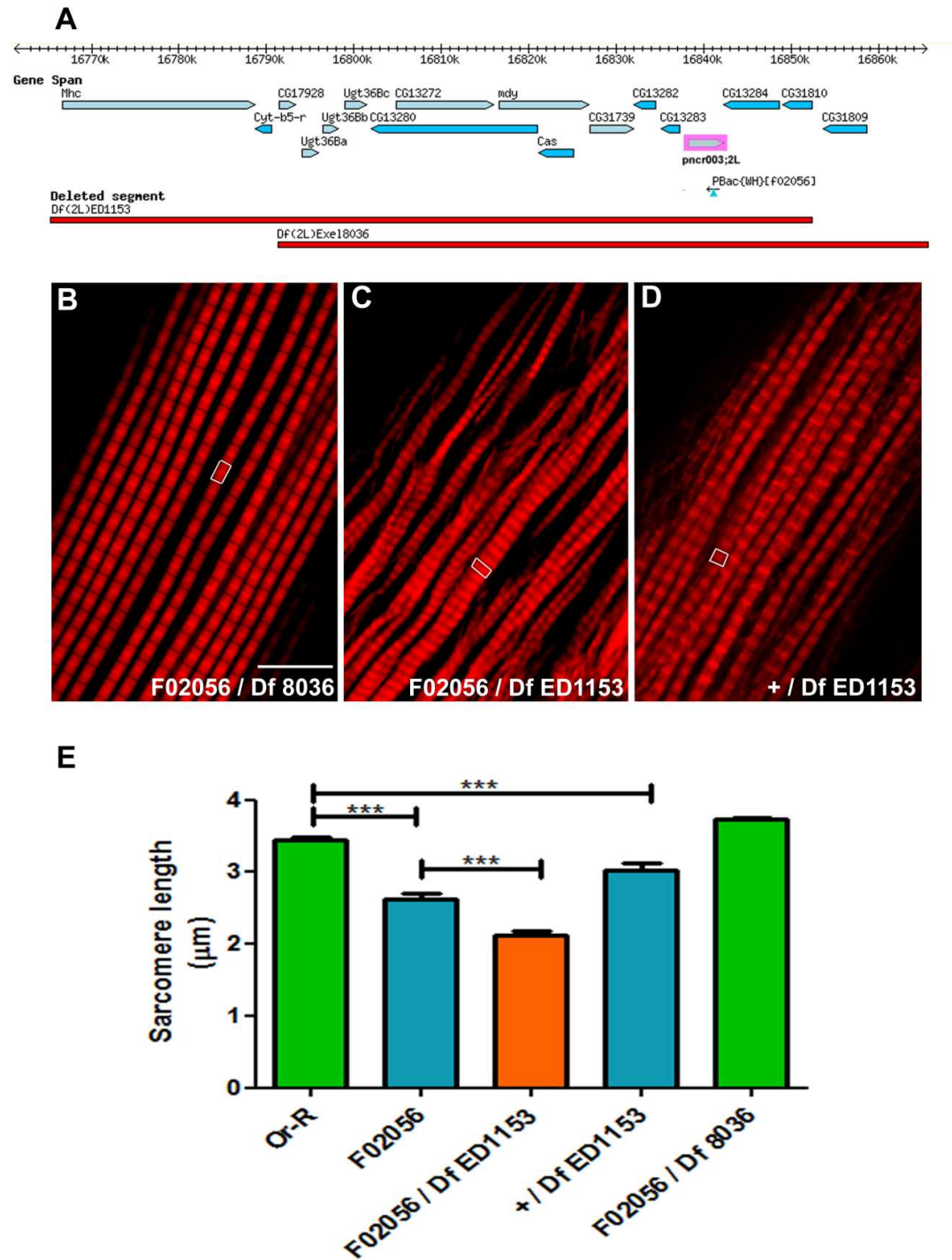


Figure 4.3

In order to confirm that *Mhc* is in fact the gene causing the phenotype, and the phenotypic enhancement observed with the deficiency *Df ED1153*, I tested whether these results could be replicated using a *Mhc* null allele (*Mhc*^{K10423}) instead of the deficiency *Df ED1153*. When heterozygous, the *F02056* condition does not show any abnormalities in sarcomere shape, or size (Figure 4.4A), indicating that the phenotype is recessive, which agrees with the lack of phenotype observed with the *Df 8036* over the *F02056* chromosome. Like heterozygous flies for *Df ED1153*, *Mhc*^{K10423} heterozygous flies also display the short sarcomere phenotype (Figure 4.4D and E), which is also enhanced over the *F02056* insertion (Figure 4.4C and E), therefore confirming that both of these effects are linked to *Mhc* itself.

2.3- Is there a strong genetic interaction between *pncr003;2L* and *Mhc*, or are these phenotypes produced by an associated *Mhc* allele?

In genetic terms, there is a clear lack of complementation between the chromosome carrying the *F02056* insertion and *Mhc*. If one were mapping the *F02056* flightless / short-sarcomere allele, without knowing which gene was affected by the insertion, these results would strongly suggest that *Mhc* itself could be the affected gene. However we know that the insertion occurred in *pncr003;2L* (Figure 4.10C provides proof of the accurate mapping of the insertion site), which is another muscle-specific gene, so it could be possible that these effects arise, instead, from a strong genetic interaction between the two genes. A genetic interaction of this kind is particularly plausible because a similar genetic interaction has already been described in the past between a dominant *Mhc* null allele and another muscle specific gene. In that case, certain *Mhc* alleles were found to suppress the phenotype of the *hdp*² allele, which has been mapped to the troponin I gene [58,59], coding for an essential component of the troponin-myosin complex, which responds to the calcium concentration in the cytoplasm to allow or inhibit the actin-myosin interaction leading to muscle contraction. A particularly interesting element of that genetic interaction, relevant to the phenomenology so far described in this manuscript, is that the *hdp*² allele is responsible for what is described as a “hypercontraction” phenotype, resulting in IFMs myofibrils with short sarcomeres [59,60]. Furthermore, because of the role of Troponin I, that phenotype is believed to arise from a misregulated response to calcium concentrations in the muscle [59]. What is interesting is that the shorter sarcomere phenotype associated with the *F02056*

insertion fits with this hypercontraction description, while the putative physiological function of *pncr003;2L* in the dyads, could also be linked to this phenotype.

Such a genetic interaction between *Mhc* and *pncr003;2L*, however, attractive as it seems, is not entirely favoured by the results described thus far in this work. Up to now, it has been shown that a lower dosage of *pncr003;2L* is not the cause of the *F02056* associated phenotype, as the deficiency *Df8036* complements the *F02056* insertion for the phenotype in question (Figure 4.3B). There seems to be a specific requirement of the *F02056* chromosome for the manifestation of the short sarcomere phenotype. This could be explained by the *F02056* insertion causing some kind of missense mutation leading to an aberrant behaviour of the peptides. However, this would only occur if the amino acid sequence of either of the peptides was compromised, and this is not the case, because the insertion occurred in the 3' UTR of the gene, and both ORF sequences are intact in the *F02056* line (see Annex 3 for sequences). The enhancement of the short sarcomere phenotype, observed when both *Mhc* and, seemingly, *pncr003;2L* are affected is also independent of the *pncr003;2L* dosage, because the *F02056* insertion in *trans* over the *Mhc* allele (+, *F02056*, / *Mhc*, +) shows the enhancement (Figure 4.4C), but not the deficiency *DfED1153* as heterozygous (Figure 4.3D), even though it completely removes both, the *Mhc* and *pncr003;2L* loci, thereby resulting in comparable haplo-dosages for both genes (*F02056,Mhc* / +, +).

The most straight forward, explanation for a complementation of the phenotype between the *F02056* insertion and the *pncr003;2L* locus (as in *Df8036/F02056*), but lack of complementation between the *F02056* insertion and the *Mhc* locus (as in *Mhc^{K1042}/F02056*), would be that the *F02056* chromosome has a defect in the *Mhc* gene itself, probably caused by a background mutation in that particular pBac line, which is responsible for the short sarcomere phenotype, and which is independent of *pncr003;2L*.

Figure 4.4: The short sarcomere phenotype enhancement is caused by a *Mhc* haploinsufficiency. (A) Diagram representing the genomic locus of *pncr003;2L* in chromosomal arm 2L, annotated as in Figure 4.1 A, and representing the putative *Mhc*^{*F02056*} allele. (B-D) Confocal microscopy images of hemithoraces stained with phalloidin-rhodamine (red), showing the myofibrils of longitudinal indirect flight muscles of (B) + / *F02056* flies, with a wild type appearance (C) *F02056* / *Mhc*^{*K10423*} flies, with an enhanced short sarcomere phenotype and (D) *Mhc*^{*K10423*} / + flies with a similar short sarcomere phenotype as that of homozygous flies for the *F02056* insertion. Representative sarcomeres are framed to highlight the differences in sarcomere shape and length between these genotypes. (E) Quantification of the differences in sarcomere length between these genotypes, showing a significant difference, as indicated by a one-tailed unpaired t-test statistical analysis, between the following genotypes: *Or-R* and *ED1153* / *F02056* flies ($t=19.59$, $p<0.0001$), *Or-R* and *ED1153* / + flies ($t=4.72$, $p=0.0005$), *ED1153* / *F02056* and *ED1153* / + flies ($t=9.55$, $p<0.0001$), *F02056* / + and *Mhc*^{*K10423*} / *F02056* flies ($t=18.02$, $p<0.0001$), *F02056* / + and *Mhc*^{*K10423*} / + flies ($t=11.97$, $p=0.0005$), *Mhc*^{*K10423*} / *F02056* and *Mhc*^{*K10423*} / + flies ($t=12.59$, $p<0.0001$). $n=200$ sarcomeres, from 4 different flies per genotype. Scale bar (B-D)=10 μ m.

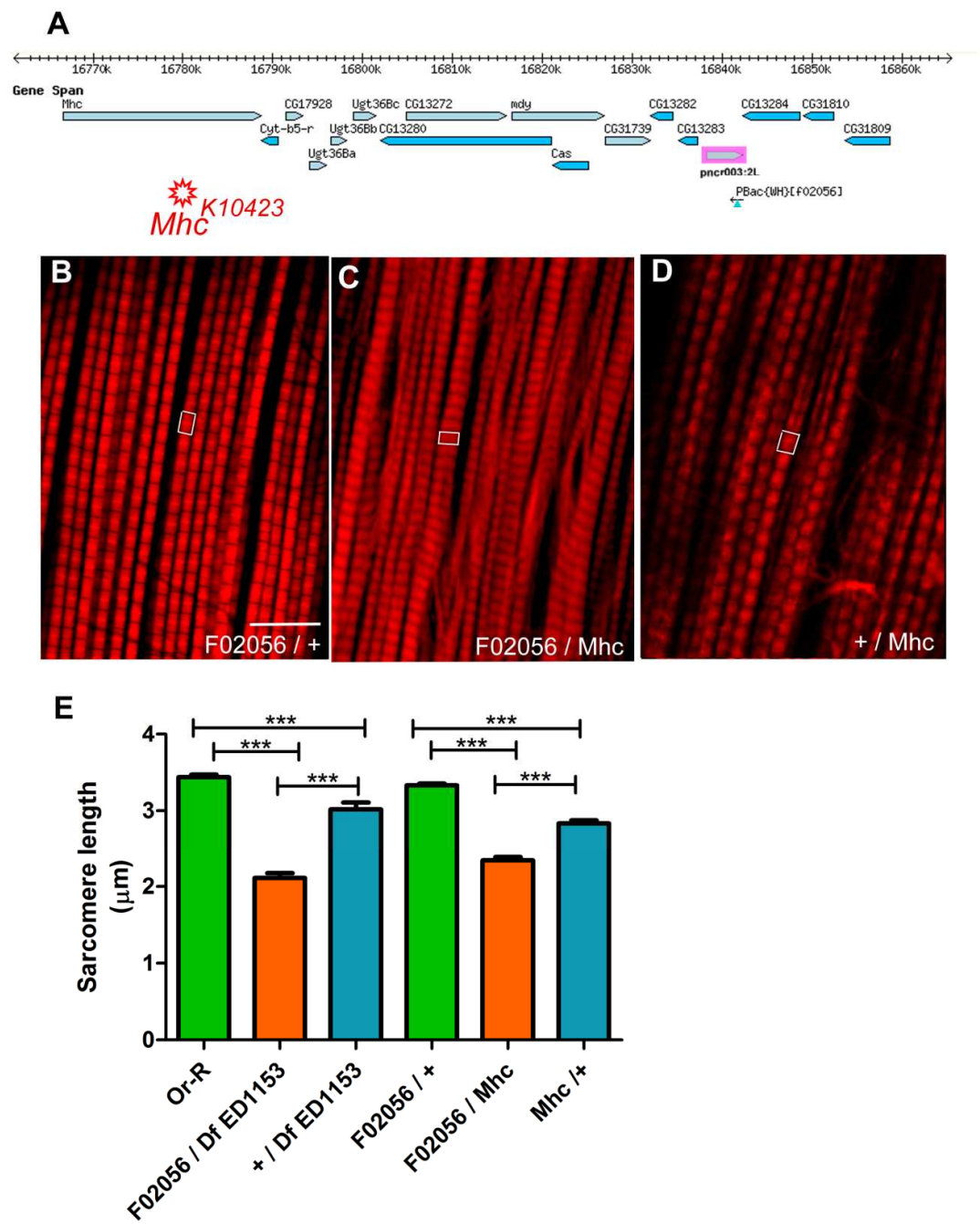


Figure 4.4

2.4- A pBac transposase-mediated reversion of F02056 restores *pncr003;2L* expression but still presents the short sarcomere phenotype.

To test whether the short sarcomere is independent of the effect of *F02056* on *pncr003;2L*, I induced the excision of the the pBac element, in order to revert the *pncr003;2L* gene to its original wild-type condition. Such a reversion, in this case, is expected to completely restore *pncr003;2L*, because *F02056* is a pBac element, known for its property to produce precise excisions in which case the sequence of their host gene is perfectly restored upon transposition of the pBac element [54].

An *F02056* revertant was obtained using a standard reversion protocol (see Annex: 2). Briefly, the expression of the pBac Transposase was brought into into the *F02056* line by means of a series of genetic crosses, and the F1 progeny was scored for the loss of the *white* marker carried by the *F02056* insertion. Although in this revertant, the sequence of *pncr003;2L* and its expression levels were restored as expected (Annex 3, Figure 4.5D), homozygous individuals still present the short sarcomere phenotype, and over the *Df ED1153* they still show a strong enhancement of the phenotype (Figure 4.5A,B and C). These results are in agreement with the existence of a background mutation in the *F02056* chromosome, independent of the *F02056* insertion itself, and responsible for the short sarcomere phenotype.

Figure 4.5: The short sarcomere phenotype is independent of the F02056 insertion.

(A) schematic representation of the RE28911 cDNA, showing the position of ORFs A and B, and the excision event of the pBac{WH}F02056 element from the 3'UTR of *pncr003;2L*. (B-C) Confocal microscopy images of hemithoraces stained with phalloidin-rhodamine (red), showing the myofibrils of longitudinal indirect flight muscles of (B) homozygous flies of the *F02056* RV2 revertant, with a similar short sarcomere phenotype as that of homozygous flies for the *F02056* insertion (C) *F02056* RV2/ *ED1153* flies, with an enhanced short sarcomere phenotype. Representative sarcomeres are framed to highlight the differences in sarcomere shape and length between these genotypes. (D) Quantification of the differences in sarcomere length between these genotypes, showing a significant difference, as indicated by a one-tailed unpaired t-test statistical analysis, between homozygous flies for the *F02056* RV2 revertant and *ED1153* / *F02056* RV2 flies ($t=5.53$, $p<0.0001$). $n=200$ sarcomeres, from 4 different flies per genotype. (E) Semi-quantitative RT-PCR on mRNA extracts of whole flies, using primers specific to either the exon 2 or exon 3 of *pncr003;2L* (using the primers represented in Figure 4.2C), showing the comparable expression of *pncr003;2L* between the *F02056* RV2 homozygous flies and wild type Or-R flies, and the visible reduction in *pncr003;2L* expression in flies homozygous for the pBac F02056 insertion. Scale bar: (B-C): 10 μ m.

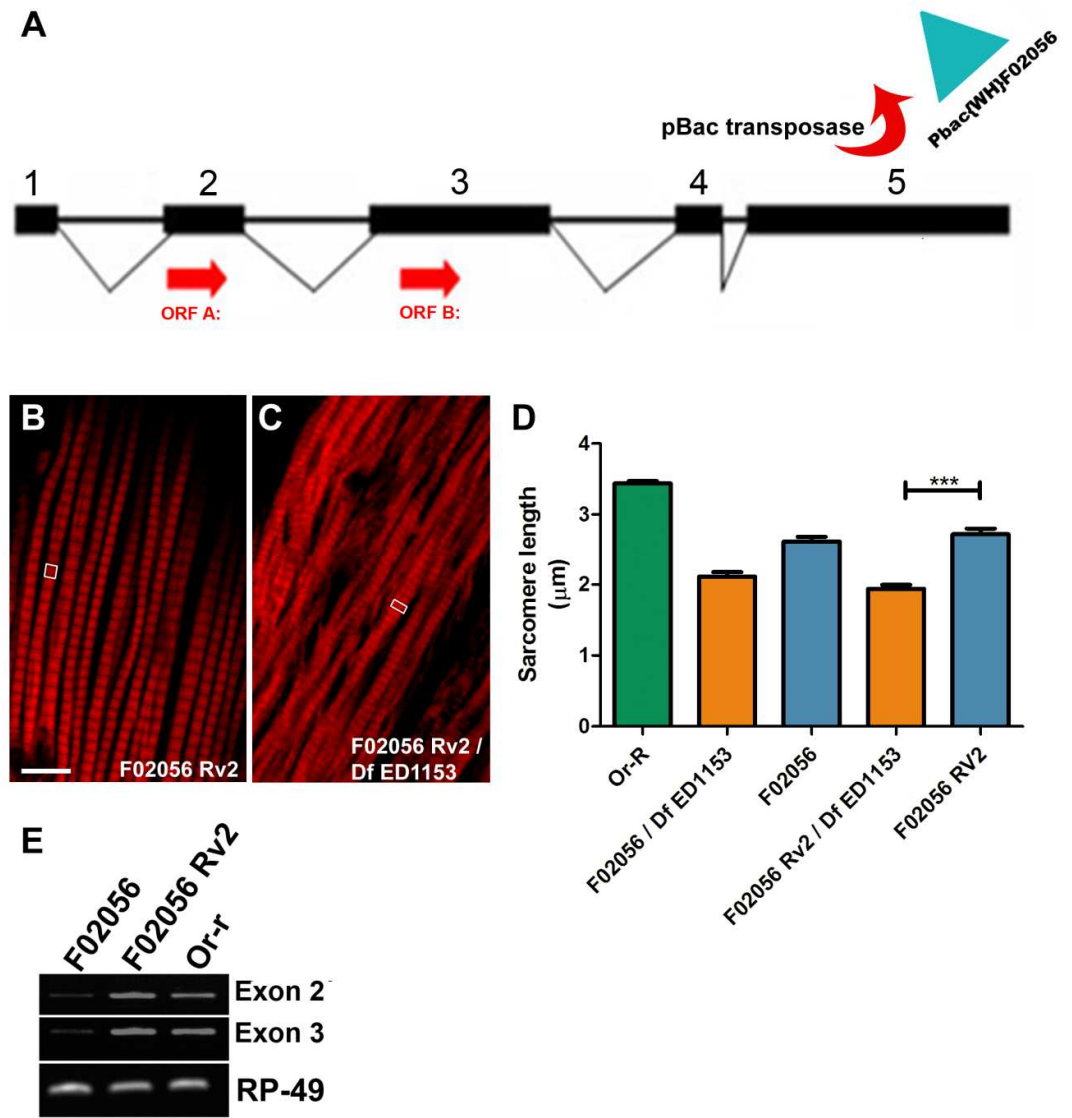


Figure 4.5

2.5- Mapping of the putative *Mhc* associated allele.

A way to test whether the background mutation maps, as hypothesised above, to the *Mhc* gene —leading to an allele which I will refer to as *Mhc*^{F02056}— is to map the allele to the *Mhc* locus by homologous recombination between the *F02056* insertion and a nearby genetic marker (Figure 4.6A). Such a genetic marker should come from a gene with a visible and easily identifiable mutant phenotype, preferably homozygous viable (in order to be able test for the homozygous *F02056* phenotype), with available alleles on the public stock centres, and located as near as possible to the *Mhc* locus. *chiffon* (*chif*), a gene coding for a zinc-finger transcription factor, involved in the development of the embryonic chorion, and cuticular structures in general [61], was identified as a good genetic marker for this approach, because it fulfilled all of these criteria.

Homozygous mutants for the *chif*^d allele have a clear “rough” eye, and scutellar bristle phenotypes, and its locus is some 400 Kb upstream of *Mhc* (Figure 4.6B). Because *pncr003;2L* is located 50 Kb downstream of *Mhc*, a chromosomal recombination between the *chif*^d and *F02056* loci, would have approximately 1 in 9 chance to occur between the *Mhc* and *pncr003;2L* loci ($450/50 = 9$), therefore, of all *chif*^d *F02056* recombinant chromosomes, about 1 / 9 should have lost the *Mhc*^{F02056} allele (Figure 4.6C). A standard homologous recombination protocol in which the F1 progeny of the cross between *w; chif^d cn¹ sca¹ bw¹ sp¹ / F02056* females and *w; chif^d cn¹ sca¹ bw¹ sp¹* males were screened for the *chif* rough eye, and thin bristle phenotypes, the retention of the *white* marker from the *F02056* insertion, and the loss of the *bw* and *sp* phenotypes. This protocol yielded 4 *chif*^d *F02056* recombinants out of 2,000 flies screened. Three out of the four recombinant lines recovered still presented the short sarcomere phenotype as homozygous for the *chif*^d *F02056* recombinant chromosome (Figure 4.7A,B,C and G), but one of them (recombinant *chif*^d *F02056* 13) had wild-type looking myofibrils, with sarcomere lengths similar to wild type (Figure 4.7D and G). Furthermore, over *Df ED1153*, the *chif*^d *F02056* 13 chromosome showed no enhancement of the short sarcomere phenotype (Figure 4.7F and H). Although the recovery of a recombinant which lacks the short sarcomere phenotype, at a rate which is not too dissimilar from that expected (1/4 compared to 1/9), is in agreement with the existence of the *Mhc*^{F02056} allele, this result only proves that the chromosomal locus of the background mutation is, as expected, downstream of *chif*, but its specific locus still needs to be determined.

Figure 4.6: Diagram representing the homologous recombination protocol

implemented to map the putative MhcF02056 allele. (A) Diagram representing the genomic locus of *pncr003;2L* in chromosomal arm 2L, annotated as in Figure 4.1 A, highlighting the lack of complementation between the putative *Mhc*^{F02056} allele and the *Mhc*^{K10423} allele, and between the *Mhc*^{F02056} allele and the *Df ED1153* covering the *Mhc* locus. (B) Broader genomic diagram of the *pncr003;2L* genomic locus, showing the genomic distance between the *Chiffon*, *Mhc*, and *pncr003;2L* genes. (C) Diagram showing the possible outcomes of the homologous recombination between *chiffon* and the *F02056* insertion. For the putative *Mhc*^{F02056} allele to be recombined out of the *F02056* chromosome, the recombination chiasm, leading to the *chif F02056* recombinant needs to occur between the *Mhc* and the *pncr003;2L* loci. This region represents about one ninth of the total length between the *chiffon* and *pncr003;2L* genes.

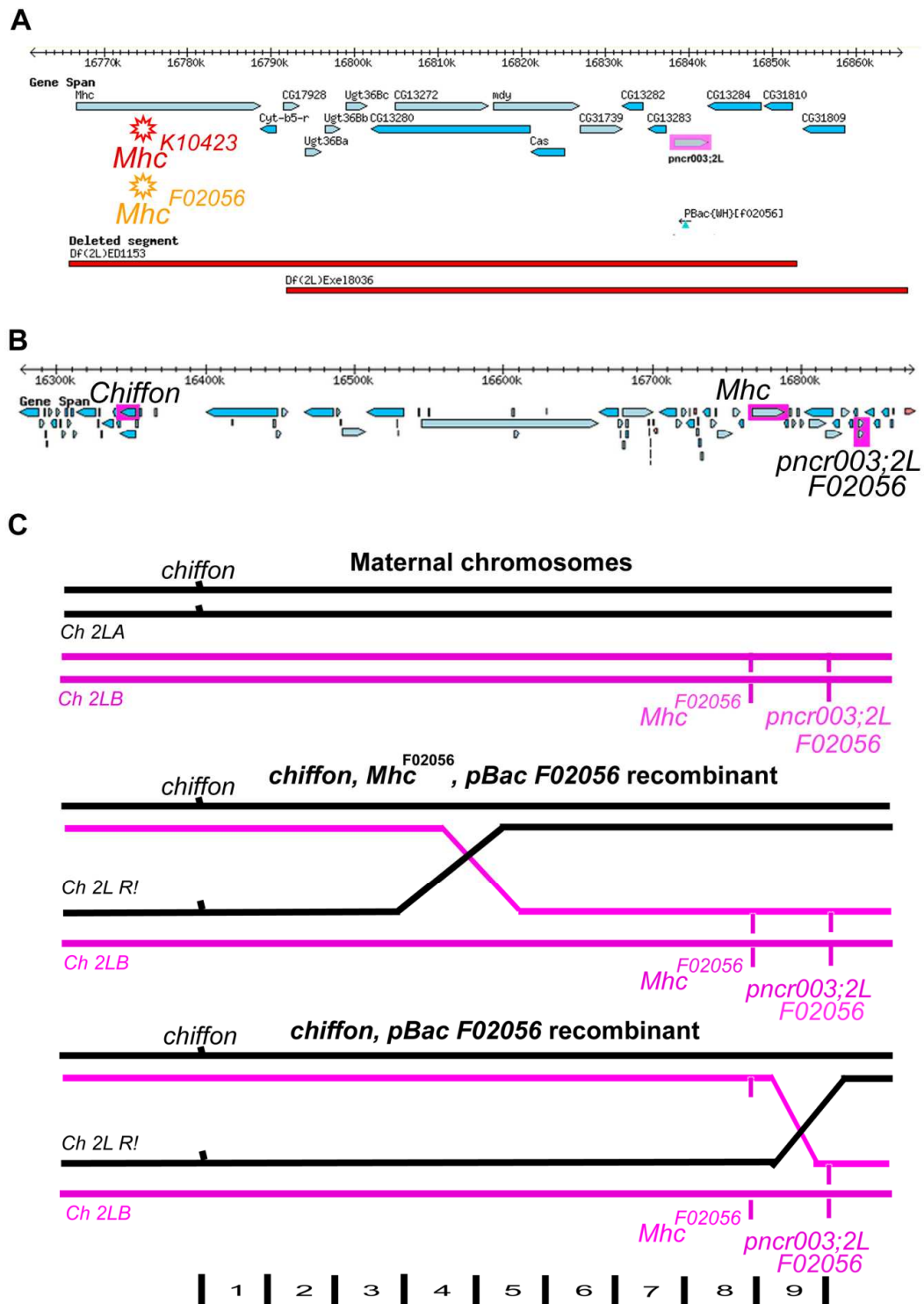


Figure 4.6

Figure 4.7: One out of four *chif* F02056 recombinants has lost the putative *Mhc*F02056 allele. (A-F) Confocal microscopy images of hemithoraces stained with phalloidin-rhodamine (red), showing the myofibrils of longitudinal indirect flight muscles of (A) *chif* F02056 recombinant 18, (B) *chif* F02056 recombinant 4, (C) *chif* F02056 recombinant 14 and (D) *chif* F02056 recombinant 13. While recombinants 18, 4 and 14 show a short sarcomere phenotype similarly to the original F02056 insertion line, the *chif* F02056 recombinant 13 line has sarcomeres with normal appearance, indicating that the recombination in this line must have occurred between the *Mhc* and *pncr003;2L* loci. In accordance with this, (E) *chif* F02056 recombinant 14 /*Df* ED1153 shows an enhancement of the short sarcomere phenotype, but (F) F02056 recombinant 13 /*Df* ED1153 does not. (G) Quantification of the differences in sarcomere length between the different *chif* F02056 recombinants, showing a significant difference, as indicated by a one-tailed paired t-test statistical analysis, between homozygous flies for the *chif* F02056 recombinant 13 and homozygous flies for the *chif* F02056 recombinant 18 ($t=5.99$, $p<0.0001$), 14 ($t=11.84$, $p<0.0001$), and 4 ($t=6.08$, $p<0.0001$). (H) Quantification of the differences in sarcomere length between the F02056 recombinant 13 /*Df* ED1153 flies and F02056 recombinant 14 /*Df* ED1153 flies, showing a significant difference, as indicated by a one-tailed unpaired t-test statistical analysis ($t=16$, $p<0.0001$). $n=200$ sarcomeres, from 4 different flies per genotype. Scale bar: 10 μm

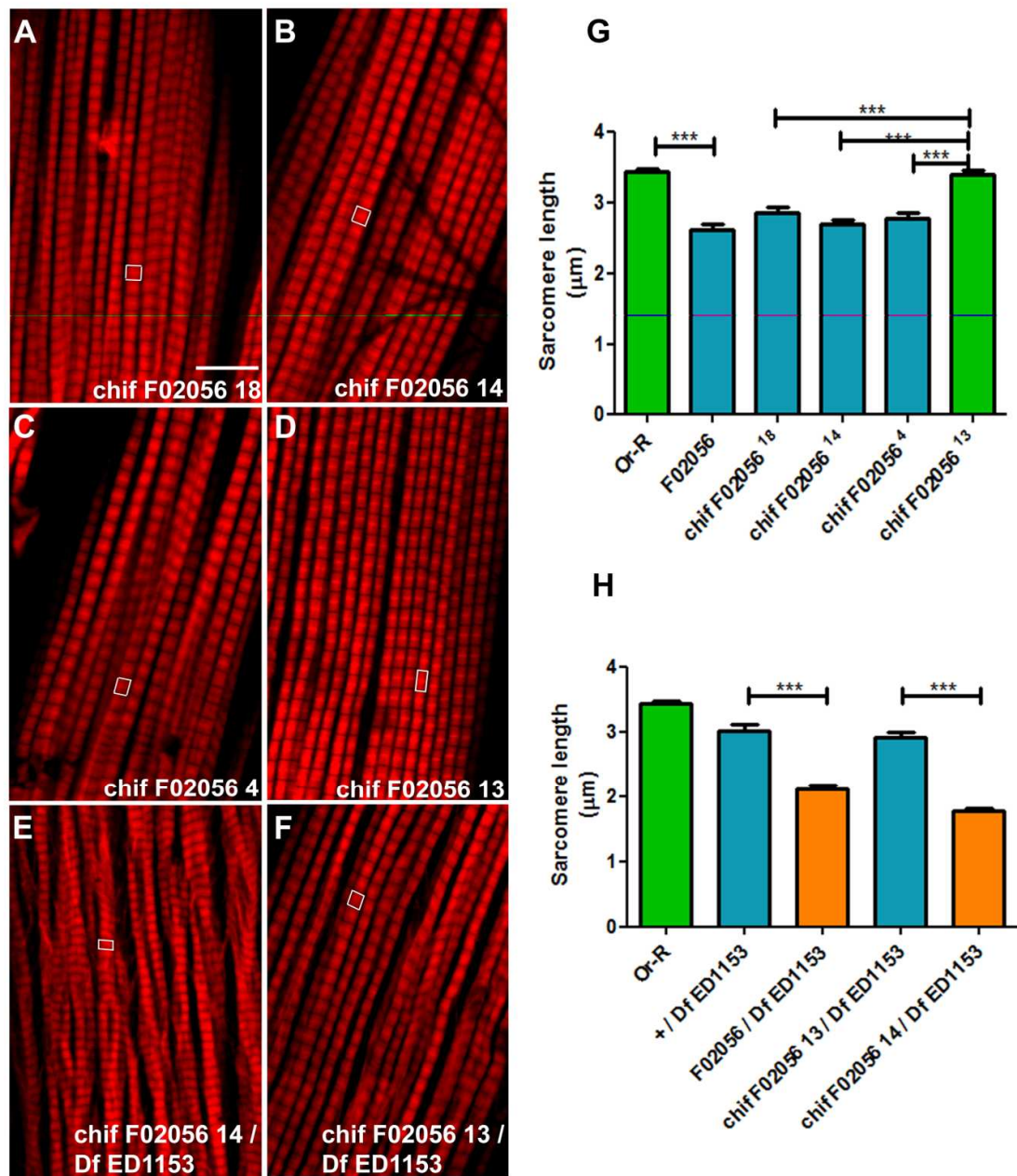


Figure 4.7

2.6- A small intronic deficiency affecting the alternative splicing of *Mhc* could be responsible for the *Mhc*^{F02056} allele.

In order to confirm that the affected gene by this background mutation is *Mhc*, I looked for possible mutations in the *Mhc* gene that could explain the observed phenotypes. Because the putative *Mhc*^{F02056} allele affects IFMs, I sequenced the genomic fragments corresponding to the exons that constitute the IFM specific *Mhc* transcript [62]. These genomic elements were amplified by PCR, using genomic DNA from wild-type and *F02056* homozygous flies, cloned and sequenced (Figure 4.8A). There are several differences between the annotated genome and these two fly lines leading to synonymous substitutions, and two mutations leading to aa substitutions, in exons 8 and 10, were detected in both wild-type and *F02056* strains; none of these differences could explain the short sarcomere phenotype (Figure 4.8B). The only seemingly relevant difference between the two strains was a 32 nt deletion in the intronic sequence immediately upstream of exon 7c.

This same region corresponding to exon 7 was amplified and sequenced in each of the four *chif* *F02056* recombinants. Interestingly, the *chif*^F *F02056*13 does not have the 32 nt deletion, whereas the other three recombinants do, indicating that the recombination that led to the loss of the short sarcomere phenotype did indeed occur, as predicted, between the *Mhc* and *pncr003;2L* loci.

Regarding the cause of the *Mhc* allele, the 32 nt deletion affecting the intronic sequence prior to the alternative exon 7c could be quite an interesting candidate since introns can contain regulatory elements which regulate the splicing of their neighbouring exons [63,64]. This intronic deletion is located very near to the exon sequence (only 13 nt upstream of it) and has sequence stretches that appear to be enriched in pyrimidine bases (Figure 4.9A). This is interesting because pyrimidine rich motifs, known as polypyrimidine tracts, are located at a similar distance from the exon, and have been shown to play an important role in the determination of both, constitutive and alternative splicing events [63,65]. Alternative splicing is particularly important for the *Drosophila Mhc* gene.

In vertebrates, there are at least 13 paralogues of the *Mhc* gene (known, in vertebrates, as *MYH*), and different tissues have been shown to express different members of the

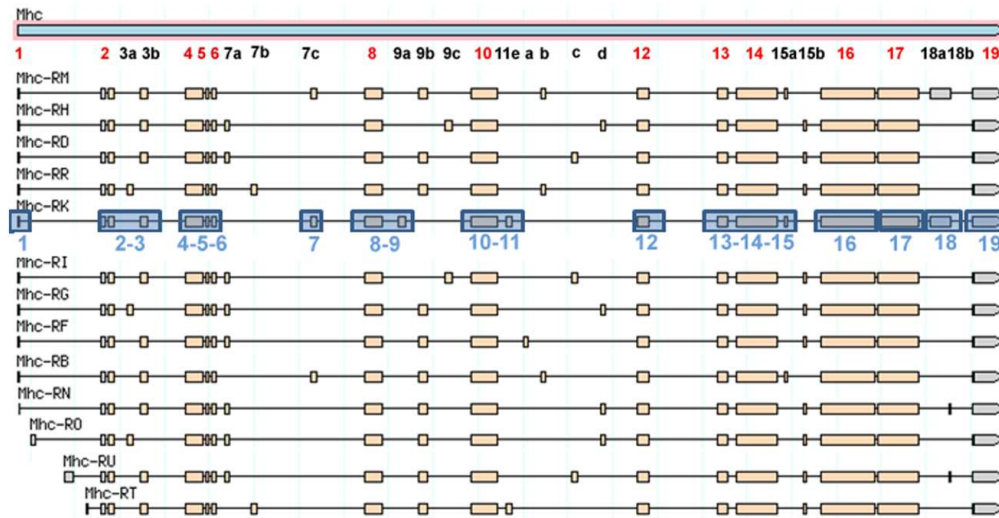
MYH gene family: there are at least 2 *MYH* genes specific to non-muscle cells, 3 *MYH* genes specific to smooth muscles, and 8 *MYH* genes to specific striated muscles [66]. The Mhc/*MYH* protein has two main functional domains: a globular head domain, which has ATPase and actin-binding activities, and performs the motor function of the protein, and a rod domain, which is necessary for the formation of the myosin filaments. Some of the different *MYH* paralogues have been shown to have different ATPase rates, which could contribute to the different contractility speeds observed in different muscles [67], therefore, the muscle specific requirement of these isoforms is important for the physiological function of the muscle.

In *Drosophila*, there is a single *Mhc* gene, but it produces several different isoforms by alternative splicing (Figure 4.8A). The *Mhc* locus consists of 13 constitutive exons and 5 groups of alternative exons, which are each composed of two to four related exons that are mutually exclusive. As in vertebrates, different *Mhc* isoforms are expressed in different kinds of muscles [62], and importantly, there is also evidence of this muscle specificity being important for muscle function. Indeed, an intronic mutation that leads to a defect in alternative splicing has already been described to generate a homozygous viable allele of *Mhc*, with the affected flies displaying defects specifically in the IFMs [68]. That particular mutation affects the splicing of exon 9a, which, interestingly, like exon 7c, encodes for a portion of the motor globular head domain of the Mhc protein.

Given the relevance of the muscle specific expression displayed by the different *MYH* paralogues in vertebrates, or by different *Mhc* isoforms in *Drosophila*, one could reason that this deficiency could give rise to the sarcomere phenotypes if it affects the splicing of the IFM specific exon 7c. To test this, I performed a semi-quantitative RT-PCR, from whole flies RNA, using primers designed to amplify the fragments corresponding to exon 7c -11e (Figure 4.9B), both preferentially included in the IFM specific *Mhc* isoform, in order to compare the presence of those particular exons between wild-type and *F02056* homozygous flies. Surprisingly, the band corresponding to the 7c -11e fragment appears visibly stronger in the *F02056* strain, while the fragments corresponding to constitutive *Mhc* exons, and to the ribosomal protein Rp-49, have comparable intensities in both strains (Figure 4.9C). This result is in line with a regulatory role of the missing intronic sequence on the alternative splicing of exon 7c, and with previous models suggesting that tissue specific binding of the Polypyrimidine tract binding protein (PTB) to the intronic region upstream of an alternative exon can

results in its exclusion [63]. As they stand, these results point to this 32 nt deletion, which affects the splicing of exon 7c as a possible cause for the myofibril phenotype observed in the *F02056* line. These results, however, are only preliminary, because to confirm all the implications stated here (such as the binding of PTB, or other factors to the intronic DNA sequence deleted in the *F02056* line) and to determine the specific effect of this misregulation would require extensive work, which, although very interesting, is beyond the scope of this thesis.

Figure 4.8: Molecular mapping of the MhcF02056 allele by sequencing the IFM specific Mhc exons. (A) Diagram representing the different alternative mRNA isoforms for the *Drosophila Mhc* gene. Constitutive exons are numbered in red and alternative exons in black. The 12 genomic fragments indicated in blue, corresponding to the exons and surrounding intronic sequences for the IFM specific *Mhc-RK* transcript [62], were amplified by PCR and sequenced. (B) Summary of the mutations found, with respect to the reference genome, by sequencing these genomic fragments. Mutations found in both wild-type and *F02056* strains are coloured in green, mutations found in only one of the two strains, but unlikely to give rise to the mutant phenotype are coloured in blue, the unique mutation that may give rise to the phenotype is coloured in red.

A**B**

| Strain sequenced: | | |
|-------------------|--|---|
| Fragment: | wild-type | F02056 |
| fragment 1 | Intron 1, 1 nuc. sub. | |
| fragment 2-3 | Intron 2-3, 1 nuc. sub.; 1 nuc. ins. | |
| | Intron 2-3, 2 nuc. sub.; 8 nuc. ins. | Intron 2-3, 2 nuc. sub.; 8 nuc. ins. |
| | Intron 2-3, 114 nuc. repeat ins. | |
| | Intron 2-3, 12 nuc. deletion | Intron 2-3, 12 nuc. deletion |
| fragment 4-5-6 | Intron 4-5, 4 nuc. subs. | Intron 4-5, 4 nuc. subs. |
| | Intron 5-6, 2 nuc. subs. | Intron 5-6, 2 nuc. subs. |
| | | Exon 6 F-285F syn. sub. |
| | | Intron 6-7 small deletion prior to exon 7 |
| fragment 7 | Exon 8, syn. aa sub. D347D | Exon 8, syn. aa sub. D347D |
| | Exon 8, aa sub. R350K | Exon 8, aa sub. R350K |
| | Exon 8, syn. aa sub. G367G | |
| | Intron 8-9, 1 nuc. sub. | Intron 8-9, 1 nuc. sub. |
| fragment 8-9 | Exon 10, aa sub. A625P | Exon 10, aa sub. A625P |
| | Intron 10-11, nuc. sub. | Intron 10-11, nuc. sub. |
| | Intron 10-11, nuc. sub. | |
| | Exon 11, syn. aa sub. A789A | Exon 11, syn. aa sub. A789A |
| fragment 10-11 | Exon 14, syn. aa sub. D943D | Exon 14, syn. aa sub. D943D |
| | Exon 14, syn. aa sub. D949D | |
| | Exon 14, syn. aa sub. L1149L * | |
| | Exon 14, syn. aa sub. S1157F * | |
| fragment 12 | Exon 14, syn. aa sub. L1177F * | |
| | Intron 14-15, 3 nuc. sub., 1 nuc. ins. | |
| | Intron 15-16, 1 nuc. sub. | Intron 15-16, 1 nuc. sub. |
| | Intron 15-16, 1 nuc. sub. | |
| fragment 13-14-15 | Exon 16 syn. aa sub. D1264 | Exon 16 syn. aa sub. D1265 |
| | Exon 16 syn. aa sub. E1387E | Exon 16 syn. aa sub. E1387E |
| | Exon 16 syn. aa sub. E1479E | |
| | Exon 16 syn. aa sub. I1515I | |
| fragment 16 | Exon 16 syn. aa sub. A1560A | |
| | Exon syn. aa sub. 17 A1677A | Exon syn. aa sub. 17 A1677A |
| | | Intron 17-18, 1 nuc. sub. |
| | | |
| fragment 17 | Exon 18, 3' UTR 1 nuc. sub. | |
| fragment fw 18 | | |
| fragment fw 19 | | |

Table key: nuc. nucleotide, aa amino acid, syn. synonymous, sub. substitution, ins. insertion.

Figure 4.8

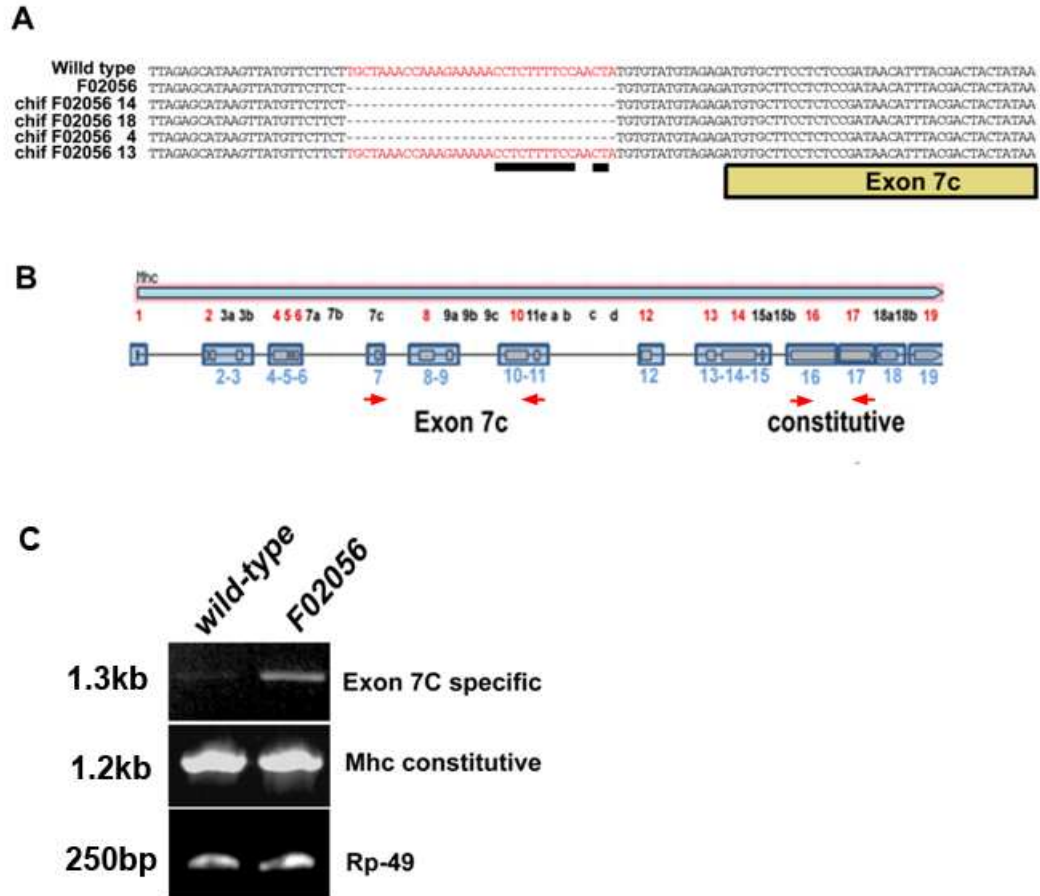


Figure 4.9

Figure 4.9: **A specific small intronic sequence deletion in *Mhc* may be the cause of the *MhcF02056* allele.** (A) DNA sequence alignment of the intronic sequence prior to exon 7c, from Or-R, the original *F02056* insertion line, and the different *chif F02056* recombinants. The small deficiency is present in the fly lines showing the short sarcomere phenotype. (B) Diagram representing the 12 genomic fragments, corresponding to the IFM specific *Mhc-RK* transcript, according to the data of Hasting *et al.* [62], which were amplified by PCR and sequenced. Red arrows represent the primers used for the exon7c specific PCR, and for the *Mhc* constitutive PCR. (C) Semi-quantitative RT-PCR on RNA extracts from whole flies, using primers to amplify specifically a fragment including exon 7c, or a fragment from constitutive exons 16 and 17.

2.7- Using the *F02056* insertion to generate a small, specific, FRT recombination-mediated deficiency.

Since the short sarcomere phenotype appears to be independent of *pncr003;2L*, the phenotype and function of this smORF gene still remains undetermined. It is possible that the reduced expression of *pncr003;2L* is still able to provide enough function to render its phenotype undetectable in the *F02056* line. Therefore it was necessary to generate a null mutant for *pncr003;2L*. Two mutagenesis strategies can be implemented, which take advantage of the *F02056* insertion to specifically remove the *pncr003;2L* locus: The first one relies on the presence of the FLP recombination target (FRT) sequences present in the Exelixis transposons, to produce an FRT mediated recombination between two transposon-carrying homologous chromosomes, leading to the deletion of the genomic sequence between the two insertions in the recombinant chromosome [56] (Figure 4.10, Annex 3). For this, it was necessary to identify an Exelixis transposon, inserted as near as possible to the *pncr003;2L* locus, in an adequate orientation to allow for the correct recombination of the FRT sites [56]. This is the case for the *pBac{RB}e01605* transposon selected for this protocol, inserted in the *CG31739* gene. The recombination between the *pBac{RB}e01605* and *pBac{WH}F02056* transposons would generate a small deletion of 10 Kb, covering *pncr003;2L* and three more genes: *CG31739*, *CG13282* and *CG13283* (Figure 3.10A). The *pBac{RB}e01605* insertion is homozygous lethal, indicating that the *CG31739* gene itself, coding for the Aspartyl/Asparaginyl-tRNA synthetase, has an essential function in the fly. The *CG13282* and *CG13283* genes encode for metabolic genes (coding for a lipase and a metalloproteinase, respectively), which have very low levels of expression according to the FlyBase expression atlas. Because the resulting recombinant transposon, in this case, would retain the *white* marker—making it virtually indistinguishable from the parental chromosomes—the standard strategy to screen for recombinants would be to generate individual lines from several F1 flies in order to screen them by PCR [56]. To facilitate this screening process, the genetic markers *black* (*b*), *cinnabar* (*cn*) and *brown* (*bw*) were recombined into the parental pBac chromosomes generating the following lines: *w; b pBac{WH}F02056 cn* and *w;pBac{RB}e01605 bw*. This way, the F1 could be screened over a *b cn bw* chromosome, and the *e01605-F02056 cn* recombinant chromosomes could be easily identified by the loss of the *b* (*black*) and *bw* (*brown*) recessive markers, which produce a distinctive darker cuticle and pale pink eyes, respectively (Figure 4.10B, Annex 3). This protocol led to the recovery of a

dozen putative recombinants. Four of these lines were tested by PCR and successfully confirmed to be deficiency-carrying recombinants (Figure 4.10C). As expected, these specific deficiencies are homozygous lethal, dying as L1 larvae, which is also the lethality phase of the *pBac{RB}e01605* homozygous animals. The *CG31739* gene would therefore have to eventually be restored, through a genomic rescue construct, in order to have a null homozygous condition for *pncr003;2L* using this small specific deletion.

Figure 4.10: Generation of an FRT-mediated specific deficiency removing the *pncr003;2L* locus. (A) Diagram representing the genomic locus of *pncr003;2L* on chromosomal arm 2L, showing the position of the two transposons, pBac{RB}e01605 and pBac{WH}F02056, used for the generation of an FRT-mediated specific genomic deletion. The region highlighted in green represents the specific genomic region removed by the deletion. (B) Diagram depicting the FRT-mediated recombination protocol leading to the generation of a specific deletion. The structures of both pBac elements are represented, showing that the FRT sites have the same orientation, and are therefore compatible with this recombination protocol. The region highlighted in green represents the specific genomic region removed by the deletion, and the chromosome highlighted in pink represents the recombinant chromosome carrying the deletion, and which has lost the *b* and *bw* genetic markers. The different genomic regions (1,2 and 3) amplified by PCR to confirm the deletions are indicated in red. Each of these regions were amplified by a combination of primers that anneal specifically to the transposable element sequence and to the genomic sequence flanking the insertion. (C) Recombinant chromosomes carrying the specific deletion yield only fragments 1 and 3 when their genomic DNA is used as template for PCR, in contrast with the parental pBac{RB}e01605 line, which only amplifies fragment 1, and the parental pBac{WH}F02056 line, which only amplifies fragments 2 and 3.

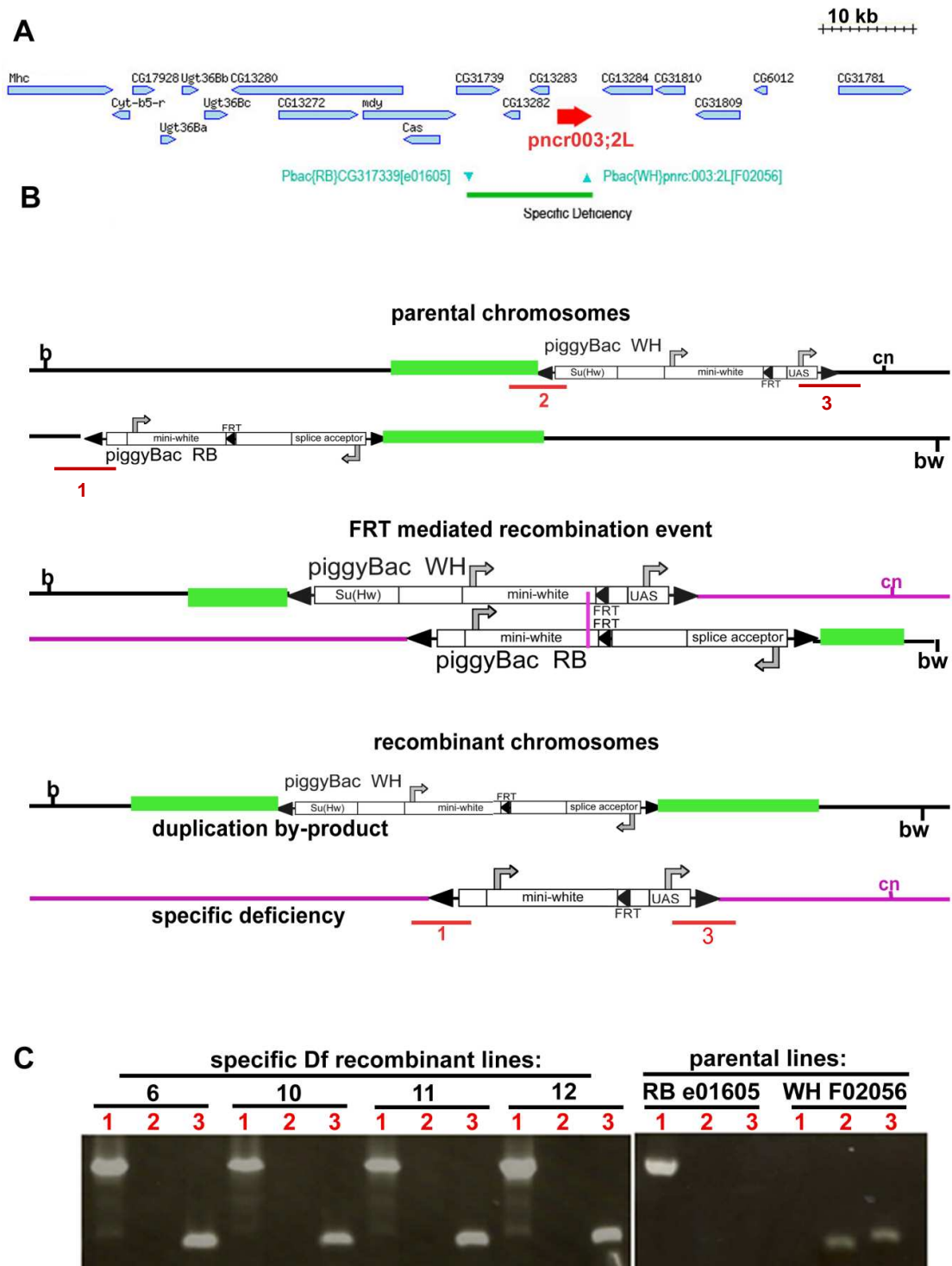


Figure 4.10

2.8- Using the *F02056* insertion to generate a directed γ -ray genomic lesion.

The second mutagenesis strategy takes advantage of the *white* marker carried by the *F02056* insertion. This method consisted of exposing young male flies, homozygous for the *F02056* insertion, to a high dose of ionizing radiation (4,500r) from a γ -ray source. This dosage of radiation has been shown to be high enough to generate double stranded DNA breaks, which can lead to genomic rearrangements, including genomic deficiencies, whilst having a minimal effect on the viability and fertility of the irradiated flies [69]. The irradiated males were crossed to *w;CyO/Gla Bc* virgin females, and the F1 progeny was scored for the loss of the white marker, which indicates that such flies have inherited a genomic lesion in the *pncr003;2L* locus (Figure 4.11). Out of 30,000 F1 progeny flies scored, 6 white eyed mutants (*Df γ -ray 1 to 6*) were recovered. Each of these mutants was mapped genetically, by crossing them to homozygous lethal alleles for several genes and deficiencies surrounding the *pncr003;2L* locus across a genomic area of some 600 kb (Figure 4.12A). This genetic mapping was complemented by a molecular confirmation of the deletion by PCR, using primers to amplify each of the boundaries of the *F02056* insertion, and in some cases, other regions of the *pncr003;2L*, and neighbouring gene (Figure 4.12B, C and D). Two of these mutants, *Dfs γ -ray 4 and 5*, are homozygous viable and complement all of the alleles and deficiencies in the region. For *Df γ -ray 4*, the PCR profile yielded both fragments, indicating that the lesion did not extend beyond the pBac element. For the *Df γ -ray 5*, only the 3' boundary of the transposon could be detected by PCR. However, it was possible to amplify a fragment within the 2nd Exon of *pncr003;2L* —using both genomic and cDNA templates— indicating that in this mutant the expression of *pncr003;2L* is not affected (Figure 4.12C). Furthermore, it was also possible to amplify a fragment corresponding to a region within the 5th exon of *pncr003;2L*, upstream of the *F02056* insertion site, indicating that in the *Df γ -ray5* line, the genomic damage did not extend much beyond the transposon in to the genomic region of *pncr003;2L* (Figure 4.12D).

The other four mutants are homozygous lethal, and of these, *Df γ -ray 1,2 and 3* fail to complement most of the genes or deficiencies in the region, indicating that these three mutants carry very large deficiencies, spanning larger regions than other already available deletions in the area (including the specific deficiencies described above, and *Df BSC325*) (Figure 4.12A). These three deficiencies also seem to cover the *Mhc* locus, and are therefore not very useful for the characterisation of a possible muscle

phenotype. The *Df* γ -ray 6 seems to complement the alleles upstream of *pncr003;2L*, including *CG31739*, and *Mhc*, but not the ones downstream, including *ApepP*, and to some extent the deficiency *DfBSC325*. Its PCR profile indicates that both boundaries of the *F02056* insertion were lost (Figure 4.12B), altogether suggesting that this mutant carries a deficiency with a breakpoint between *CG31739* and *pncr003;2L*. Most interestingly, it was not possible to amplify, by PCR, a fragment corresponding to the 2nd exon of *pncr003;2L*, in flies carrying this new deletion *Df* γ -ray 6 over the deficiency *Df* *ED1153*, which completely covers the *pncr003;2L* locus (Figure 4.12C). This indicates that the *Df* γ -ray 6 also covers most, if not all of the *pncr003;2L* locus. The fact that this deletion is viable over *Df* *ED1153*, and over the *CG31739* lethal allele, while covering most of *pncr003;2L* is extremely interesting, and even fortunate, because it should then also be viable over the specific deficiency described above, giving rise to a *pncr003;2L* null condition, which by-passes the need for a genomic rescue for the *CG31739* gene discussed above.

Crossing the deficiency *Df* γ -ray 6 over a specific deficiency (*Df*(12)) does indeed give rise to flies that are viable as adults, and which do not show any immediately visible defects. In these flies, it was not possible to amplify a genomic PCR fragment for *pncr003;2L*, nor *CG13282*, but it was possible to amplify a fragment corresponding to *CG31739* (Figure 4.13B), showing that one of the breakpoints of the *Df* γ -ray 6 deletion has to be located between the two genes *CG13282* and *CG31739* (Figure 4.13A). Although the other breakpoint of the *Df* γ -ray 6 has not been molecularly mapped, the data from the genetic complementation experiments indicates that it should be located downstream of the *ApepP* gene and upstream of, or near to, the breakpoint of the deficiency *Df* *ED1156* (See Figure 4.12A), hence, this deficiency completely covers the *pncr003;2L* locus. When trans-heterozygous, the overlapping area between *Df* γ -ray 6 and *Df*(12) produces a synthetic deficiency, which will be referred to as *Df* *pncr003;2L* (Figure 4.13 A), and which represents a *pncr003;2L* null condition. In agreement with this, *in situ* hybridisation with a *pncr003;2L* specific probe shows no expression of *pncr003;2L* in *Df* *pncr003;2L* larvae (Figure 4.13 C-D) nor adults (Figure 4.13 E-F).

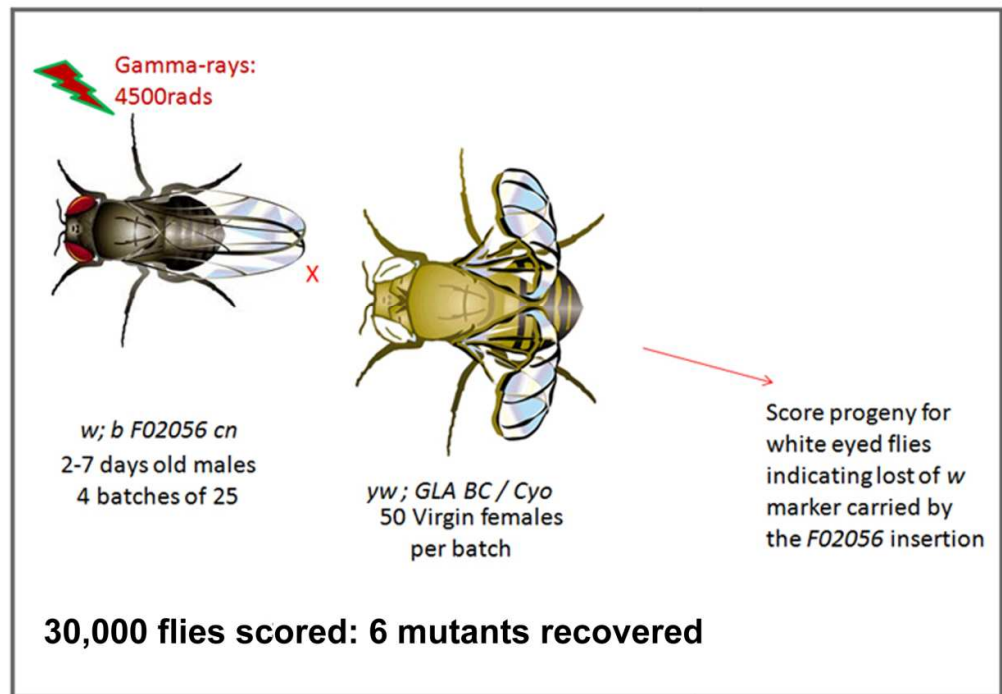


Figure 4.11.

Figure 4.11: Generation of a Gamma-ray induce a deletion targeting the *pncr003;2L* locus. Diagram representing the protocol followed to generate the gamma-ray genomic lesions targeting the *pncr003;2L* locus.

Figure 4.12: genetic and molecular mapping of the Gamma-ray mutants targeting the *pncr003;2L* locus. (A) Diagram representing a 500Kb genomic region of chromosomal arm 2L, surrounding *pncr003;2L*, showing the locus of the homozygous null alleles *PBac{RB}CG42389^{e02963}*, *Mi{MIC}ApepP^{MI01970}*, *Mi{ET1}mdy^{MB08748}*, *P{lacW}Mhc^{k10423}*, and *pBac{RB}CG31739^{e01605}*, affecting the highlighted genes: *CG42383*, *ApepP*, *midway(mdy)*, *Mhc*, and *Cg31739*, respectively, and the genomic deficiencies (red lines) used for the genetic complementation assay to genetically map the γ -ray mutants. Homozygous viable γ -ray mutant lines highlighted in green are homozygous viable, and in red are homozygous lethal. Solid and open circles represent complementation results with alleles and deficiencies, respectively. Green circles indicate complementation, and black circles lack of complementation. (B) PCR assay of the presence of the *pBac{WH}F02056* transposon sequence using genomic DNA extracts of the indicated lines. The fragments amplified by the PCR reactions are indicated by the red lines, labelled as “a” and “b”, in the diagram. Each of these regions was amplified by a combination of primers which anneal specifically to the transposable element sequence and to the genomic sequence flanking the insertion. (C) PCR assay of the presence of the *pncr003;2L* sequence, using either genomic DNA extracts, or cDNA from RNA extracts, for each of the indicated genetic conditions. (D) PCR assay to determine the extent of the γ -ray 5 genomic lesion, using genomic DNA extracts of the γ -ray 5 mutant line, or the F02056 insertion line. The “a” and “b” fragments are the same as those indicated in (B), and the Exon 5 region is indicated in the top diagram of this panel.

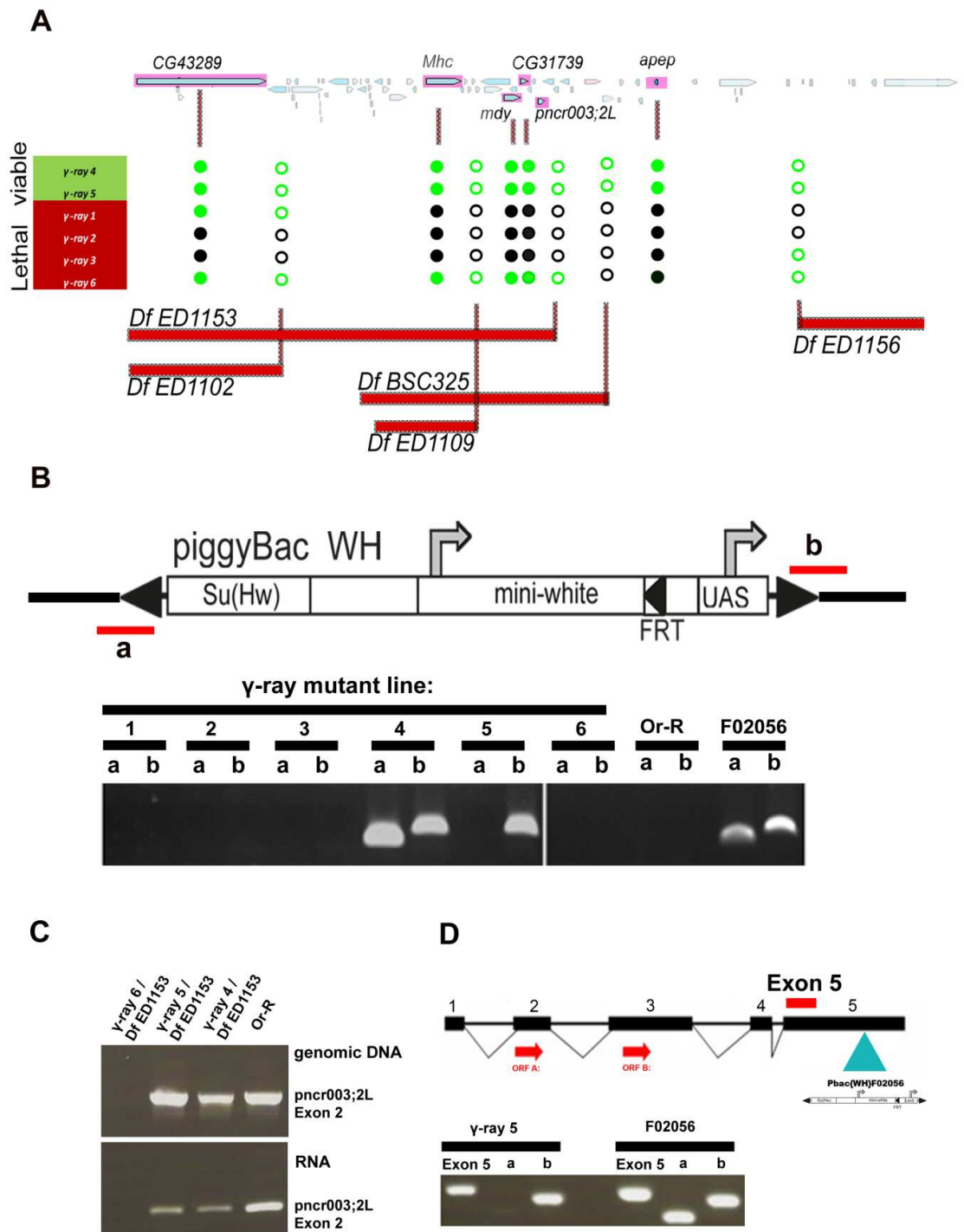


Figure 4.12

Figure 4.13 When trans-heterozygous, the deficiencies *Df γ-ray 6* and *Df 12*, give rise to a homozygous viable null condition for *pncr003;2L*: (A) Diagram representing the molecular mapping of the *γ-ray 6* breakpoint. The genomic regions amplified by the PCR reactions shown in (B) are indicated by open red squares (labelled 1 to 3) above their respective genes. The regions removed by the deficiencies *Df ED1153*, the specific *Df(12)*, and *Df γ-ray 6* are indicated. The grey area within *Df γ-ray 6* represents the possible region where the *Df γ-ray 6* breakpoint could be located. The overlap between the regions removed by *Df γ-ray 6* and *Df(12)* in flies trans-heterozygous for these two deficiencies, represents the synthetic deficiency *Df pncr003;2L*, which completely removes *pncr003;2L*. (B) PCR assay to map the breakpoint of *Df γ-ray 6*, using genomic DNA from the indicated genotypes, and primers to amplify the three regions (1, 2 and 3) indicated in (A), showing that *Df γ-ray 6* removes *pncr003;2L* and the 5' region of *CG13282*, but does not remove the 3' region of *CG31739*. and that over *Df(12)* it gives rise to a null condition for *pncr003;2L*. (C-F) *in situ* hybridisation experiments, using probes specific to the *pncr003;2L* transcript, performed in (C-D) L3 larvae and (E-F) adult abdomens. The strong expression of *pncr003;2L* observed in (D) the somatic muscles (arrow heads) and heart (arrows) of wild-type larvae and in (F) the wild-type adult heart (arrow heads), is completely lost in *Df pncr003;2L* tissues (C and E). Scale bars: 500µm.

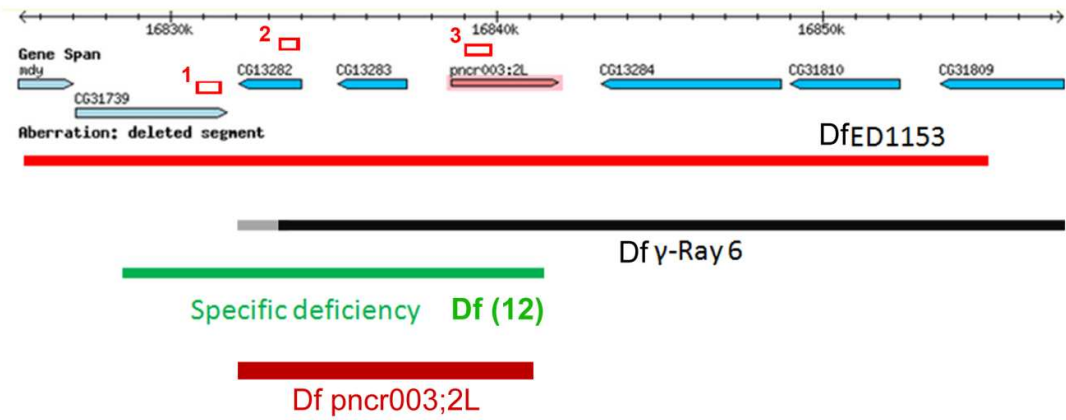
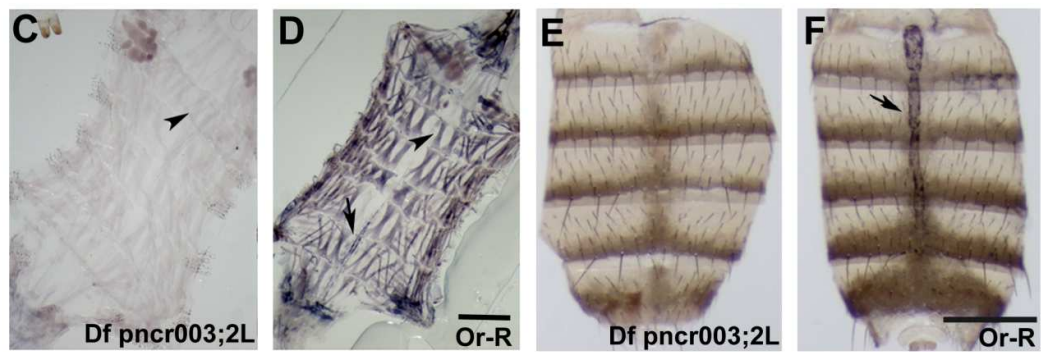
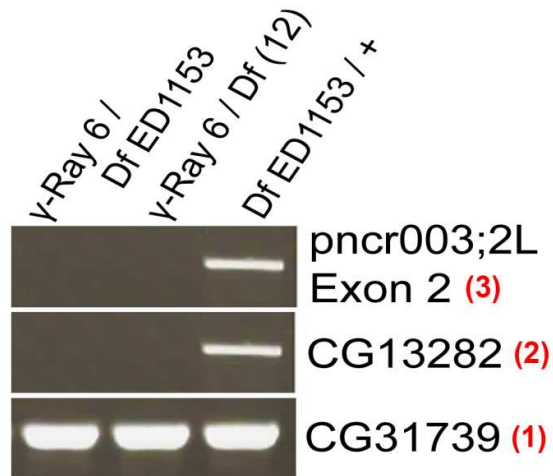
A**B**

Figure 4.13

These *pncr003;2L* null flies show no visible phenotype, and their IFM myofibrils have a wild-type appearance, with their sarcomeres measuring the average size of 3.4µm, showing that *pncr003;2L* is not responsible for any structural or morphological phenotype in these muscles (Figure 4.14 A and D). In order to assess their structure with a higher resolution, I used transmission electron micrographs to analyse the *pncr003;2L* myofibrils (Figure 4.14 E-H). No defects could be detected in neither transversal, nor longitudinal sections of the myofibrils. Importantly, the dyads, which are the particular structures where *pncr003;2L* peptides localise, also appeared to have a normal morphology and localisation.

While these results could point to some sort of physiological role for *pncr003;2L*, they are also in accordance with the above statement that it is the *Mhc*^{F02056} allele, and not *pncr003;2L*, that is responsible for the short sarcomere phenotype. The deficiency *Df*(12) over a *Mhc* null allele does not show any enhancement of the short sarcomere phenotype, but the deficiency *Df*γ-ray 6 does (Figure 4.14 B, C and D), this is as expected since *Df*(12) comes from a recombination event, in which the *Mhc*^{F02056} allele was lost, but *Df*γ-ray 6 still carries the same *Mhc* locus as the *F02056* line.

These *pncr003;2L* mutants do not appear to have a motility phenotype, in particular, their flight capabilities, which often correlate well with mutations affecting muscle function [70,71,72], seem to be normal in these mutants. This was confirmed using a method slightly more sophisticated than the test previously described, consisting of releasing the flies at the top of a large oil coated cylinder and recording at which height in the cylinder they landed [70]. Wild-type and *Df pncr003;2L* flies perform similarly in this test, with most flies getting stuck to the top two sections of the cylinder (Figure 4.15 A and B). On the other hand flies homozygous for the *F02056* insertion have a more homogenous distribution across the cylinder, with a higher number of flies falling at the very bottom of it (Figure 4.15 C), although these flies show a flight defect, this is not as severe as the “flightless” *Mhc* null condition, in which most flies fall at the bottom (Figure 4.15 D), and very few reach the top sections. These results are in agreement with the preliminary observations indicating that no motility issues exist in these *pncr003;2L* mutants, and confirm the slight flight defect observed for the *F02056* line (Table 4.1). It is important to note that all of these results indicate that the *Mhc*^{F02056} allele is recessive, as opposed to a *Mhc* null allele, since no morphology or behavioural abnormalities associated with its homozygous condition were detected in the *Df*

pncr003;2L mutants, nor in *F02056* heterozygous flies. This is important as any phenotype observed for the *Df pncr003;2L* condition in a more specialised muscle function study is likely to be associated with the *pncr003;2L* gene itself.

Figure 4.14: *pncr003;2L* null flies have no structural abnormalities in their muscle organisation. (A-C) Confocal microscopy images of hemithoraces stained with phalloidin-rhodamine (red), showing the myofibrils of longitudinal indirect flight muscles of (A) *Df* γ -ray 6 / *Df*(12) flies (*Df pncr003;2L* mutants), with wild-type looking sarcomeres, (B) *Df*(12)/ *Mhc*^{K10423} flies, with a similar short sarcomere phenotype as that of haploinsufficient flies for *Mhc* and homozygous flies for the *F02056* insertion, and (C) *Mhc*^{K10423} / *Df* γ -ray 6 flies, showing an enhancement of the short sarcomere phenotype. Scale bar: 10 μ m (D) Quantification of the differences in sarcomere length, showing a significant difference, as indicated by a one-tailed paired t-test statistical analysis, between the *Df*(12)/ *Mhc*^{K10423} and *Mhc*^{K10423} / *Df* γ -ray 6 conditions ($t=3.972$, $p<0.0032$), and the *Df* γ -ray 6 / *Df*(12) and *Df*(12) / *Mhc*^{K10423} ($t=5.128$, $p<0.0002$). $n=200$ sarcomeres, from 4 different flies per genotype. (E-F) Transmission Electron Microscopy (TEM) micrographs of longitudinal sections of adult IFMs from (E) wild-type flies, and (F) *Df pncr003;2L* mutants, in both cases, regularly organised rows of sarcomeres, displaying similarly spaced Z and M bands, can be observed, as well as the dyads (arrows) and mitochondria (dark grey rounded structures) abutting the sarcomeres. These images reveal that no ultra structural defects exist in the sarcomeric structure and dyads (arrows) of *Df pncr003;2L* mutants. Scale bars: 1.5 μ m. (G-H) TEM micrograph of transversal sections of indirect flight muscles in (G) wild-type flies, and (H) *Df pncr003;2L* mutants. In both cases the myofibrils appear as round structures showing a regular pattern of thin (actin) and thick (myosin) filaments, and with the dyads in close apposition to the sarcomeres (arrow), indicating that no structural abnormalities exist in the sarcomeres nor in the dyads of *pncr003;2L* muscles (arrow). Scale bars: 0.5 μ m.

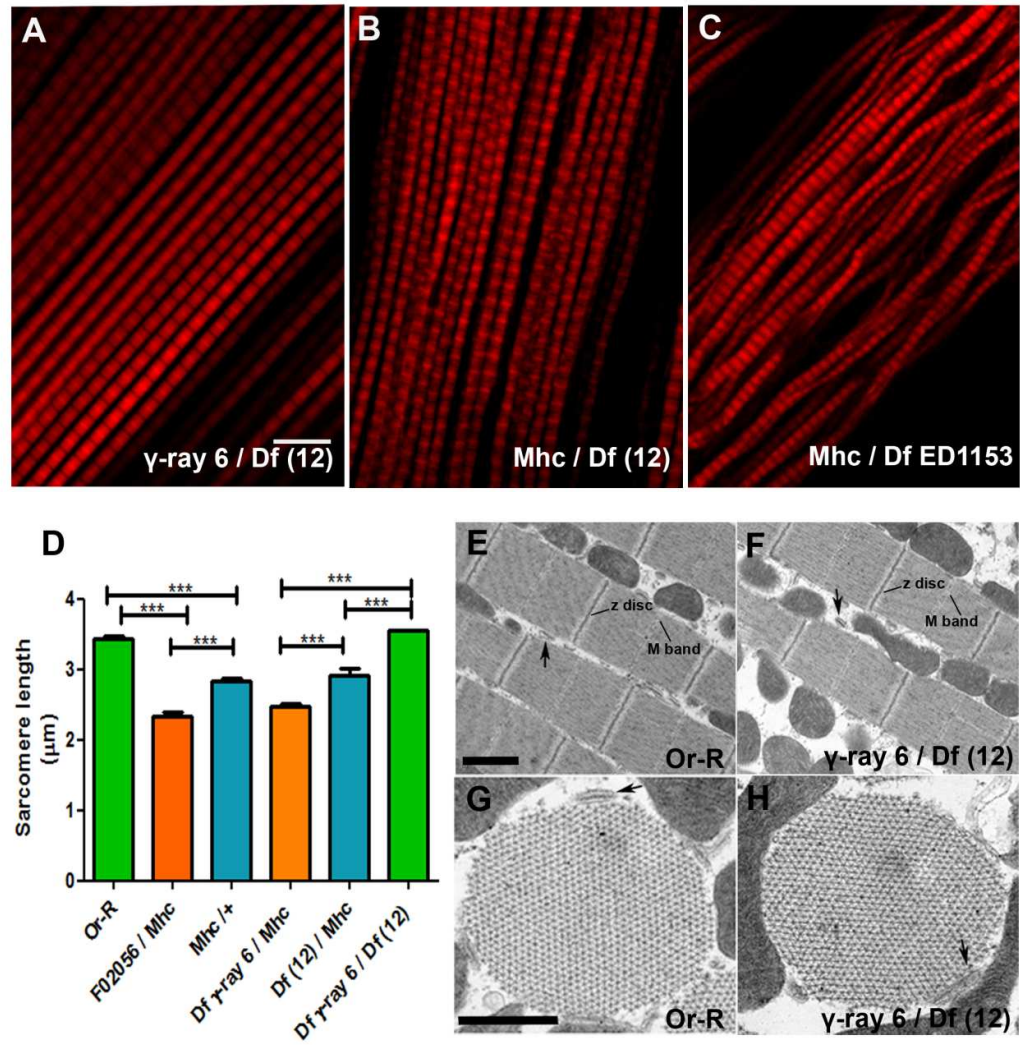


Figure 4.14

Figure 4.15 Quantitative flight assays indicate that *pncr003;2L* null flies can fly normally: (A-D) Horizontal bar charts representing the results of the flight assays (see methods, section 2.15) [70]. (A) *Or-R* flies, and (B) *Df pncr003;2L* flies perform equally well in this test, with most of the flies flying to the top of the cylinder. (C) *F02056* homozygous flies have a more spread distribution across the cylinder, indicating a slight flight defect for these flies, compared to *Or-R* and *Df pncr003;2L* flies. (D) Flightless flies heterozygous for a *Mhc* null condition, fall in their great majority to the bottom of the cylinder.

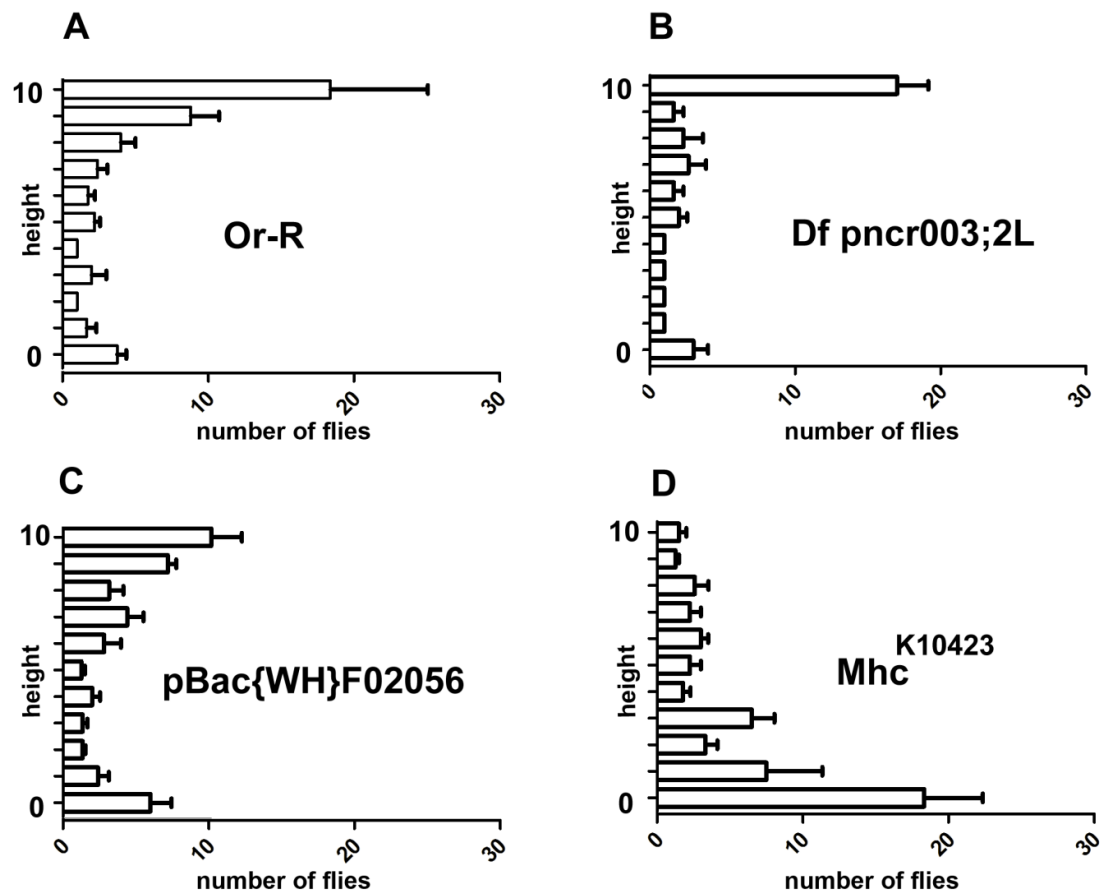


Figure 4.15

2.9- Generation of *pncr003;2L* null backgrounds free of the *Mhc*^{F02056} allele.

Although the *Df pncr003;2L* flies show no behavioural or morphological phenotype, in IFMs, which indicated that the *Mhc*^{F02056} is recessive these tissues, and most likely specific to indirect flight muscles, it may be necessary to rule out a possible implication of the *Mhc*^{F02056} allele in other tissues. This is particularly the case, as the semi-quantitative RT-PCR for the IFM specific exon 7c in whole flies, suggests that this exon may be expressed at higher levels than wild-type, and therefore the possibility exists that it may be misregulated in other tissues.

Two strategies were implemented to remove the *Mhc*^{F02056} allele from the *pncr003;2L* null background. The first strategy involves the genomic rescue of the *CG31739* gene, in the *Df(2L)12* background in order to allow this deficiency to be homozygous viable, while remaining a null for *pncr003;2L*. For this genomic rescue, a genomic sequence of 6,696 bp, including the *CG31739* genomic sequence, and the upstream and downstream genomic regions between *CG31739* and its neighbouring genes (*Cas*, and *CG13282*) — included in an attempt to preserve the important regulatory regions of *CG31739*—, was amplified by PCR, sequenced, and cloned into the *Drosophila* transgenesis vector pCaSpeR 5, which was then used to generate transgenic flies (Figure 4.16).

Homozygous viable insertions in chromosome II were selected, and linked by homologous recombination to the *Df(2L)12* chromosome. Because both the genomic *CG31739* (*gCG31739*) insertion and *Df(2L)12* carry a *white* gene marker, it was possible to select for recombinants by screening for F1 males from the progeny of *gCG31739/ Df(2L)12* females and *CyO/If* males, with dark red eyes. These *gCG31739, Df(2L)12* recombinant chromosomes are homozygous viable, showing that the construct is able to rescue the lethality of the *CG31739* null condition, and the flies carrying them, like the *Df pncr003;2L* flies, do not present any visible abnormal phenotypes.

The second strategy used another transposon-based method, known as P-element induced male homologous recombination [73]. This method takes advantage of a well described phenomenon associated with the transposition event of P-elements, in which the recombination of homologous chromosomes in males, which does not usually occur in *Drosophila*, takes place in a site specific way at the ends of the P-element insertion.

Because this method leads to homologous recombination in a site-specific manner, it has been widely used to map mutations in the *Drosophila* genome. In this case, it can be used to generate a recombinant between the *Df* γ -ray 6 deficiency, and a transposon inserted between *Mhc* and the *Df* γ -ray 6, replacing the *Mhc*^{F02056} allele with a wild-type *Mhc* locus. This approach is very similar to that described above with the *chiffon* allele, but it offers the advantage of being site-directed, and therefore, there is no need to screen for positive recombinants by analysing their IFM phenotypes, while taking the “chance” factor out of the protocol. The transposon insertion used for this method was the homozygous viable, *white* marker-carrying, *P{SUPor-P}KG07247* P-element (Figure 4.17 A), which is inserted in the 5'UTR of the *CG17928* gene —encoding for a fatty-acid desaturase, with low to moderate levels of expression. For this protocol the F1 progeny of the cross between males carrying the *P{SUPor-P}KG07247* insertion over the *b Df* γ -ray 6 *sp* chromosome and the Δ 2-3 transposase (*w*; *P{SUPor-P}KG07247* / *b Df* γ -ray 6 *sp*; Δ 2-3 /+), and females carrying the selection chromosome (*w* ; *al b sp*), were screened for recombinants having lost the *b* marker, gained the *white*+ marker, and retained the *sp* marker (Figure 4.17 B, Table 4.2). Two *w*; *P{SUPor-P}KG07247*, *Df* γ -ray 6 *sp* recombinants were recovered out of 3,120 flies scored in this screen, in which the genotypes of excision, and re-insertion events of the P-element were also observed (Table 4.2). These male recombinants are homozygous lethal, in accordance with the presence of the *Df* γ -ray 6 deficiency, and over the *Df* *ED1153*, which entirely removes the *Mhc* locus, the recombinant chromosomes do not show the enhancement of the short sarcomere phenotype (Figure 4.17 C). In accordance with this, sequencing the intronic sequence upstream of exon 7c in these male recombinant flies shows that the deficiency associated with the *Mhc*^{F02056} allele is no longer there (Figure 4.17 D). Like with the original *Df* γ -ray 6 chromosome, flies carrying the male recombinant chromosome (*P{SUPor-P}KG07247*, *Df* γ -ray 6 *sp*) over *Df*(2L)12, show no visible abnormalities.

Figure 4.16: Diagram representing the method followed to rescue the *CG31739* gene, in the *Df(2L)12* background. The genomic sequence of 6,696 bp, including the *CG31739* gene sequence, and the upstream and downstream genomic regions between *CG31739* and its neighbouring genes (*Cas*, and *CG13282*) was amplified by PCR, using a two step method, by amplifying two contiguous PCR fragments, annotated as 1 and 2, of 3Kb and 3.7 Kb, which were then ligated together by the addition of the unique *NheI* restriction site at the 3' of the left fragment (1) and the 5' of the right fragment (2). These two fragments were sequentially cloned into the pCaSpeR5*Drosophila* transgenesis vector, which was used to generate transgenic flies. Homozygous viable insertions in chromosome II were selected, and linked by homologous recombination to the *Df(2L)12* chromosome.

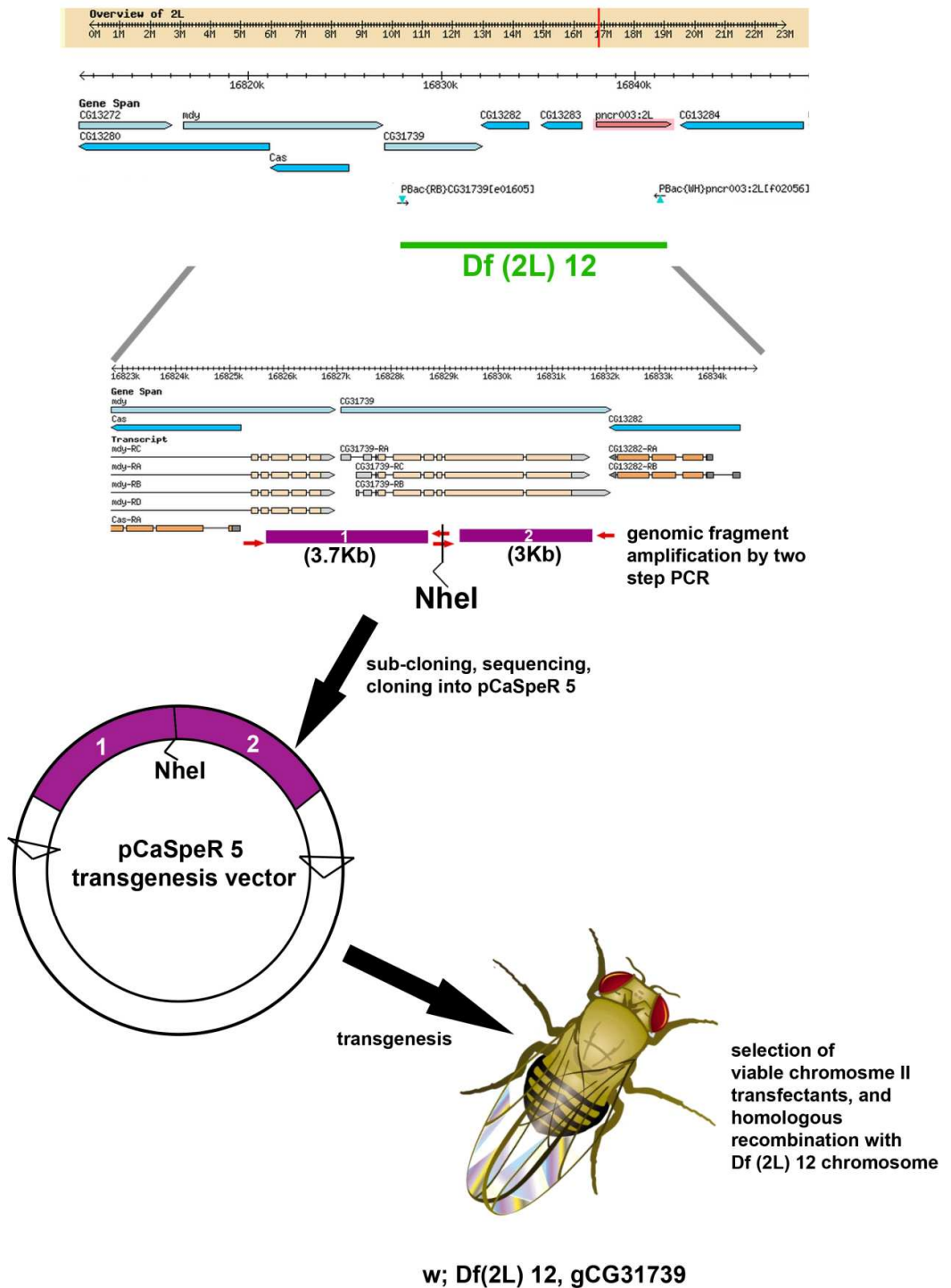


Figure 4.16

Figure 4.17: Male recombination protocol to remove the MhcF02056 allele from the Df γ -ray 6 genetic background. (A) Diagram of the genomic locus of the insertion *P{SUPor-P}KG07247* (*KG07247*), inserted between *Mhc* and *pncr003;2L*, in the 5'UTR of the *CG17928* gene chosen for the male recombination protocol. (B) Schematic representation of the $\Delta 2$ -3 transposase-mediated recombination event, which can take place in *w; KG07247 / b Df γ -ray 6 sp; $\Delta 2$ -3 / +* males. Male recombinant chromosomes can be screened by the loss of the *b* marker, the gain of the *white+* marker, and retention of the *sp* marker. (C) The male recombinant chromosome (*KG07247, Df γ -ray 6 sp*) has lost the ability to enhance the *Mhc*-related, short sarcomere phenotype, as indicated by the significant difference in sarcomere length, between the *KG07247, Df γ -ray 6 sp / ED1153* and *b Df γ -ray 6 sp / ED1153* conditions, as indicated by a one-tailed paired t-test statistical analysis ($t=16.7$, $p<0.0001$). $n=200$ sarcomeres, from 4 different flies per genotype. (D) Alignment of the DNA sequences corresponding to the intronic sequence prior to exon7c of *Mhc*, from Or-R, *KG07247, Df γ -ray 6 sp* male recombinants, and the original *b Df γ -ray 6 sp* chromosome, showing that the male recombinant has lost the small intronic deficiency.

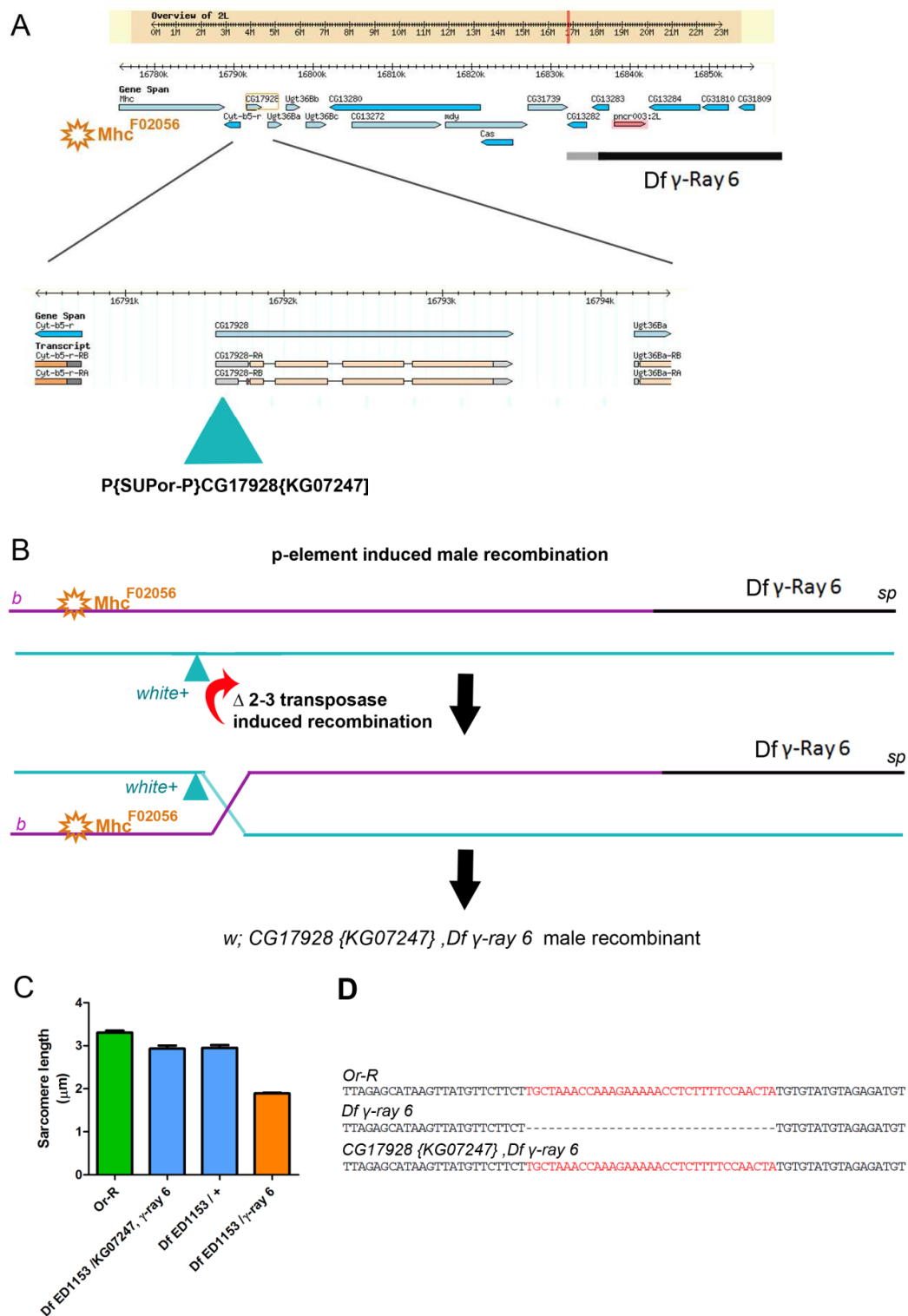


Figure 4.17

| Class | Genotype | phenotype | n |
|-------------------------|---|-------------------------|----------|
| parental | <i>b Mhc[F02056], Df γ-ray 6 sp /al b sp</i> | white eye, black, speck | 1459 |
| parental | <i>KG07247 (w+) /al b sp</i> | red eye | 1375 |
| excision | <i>KG07247 (w+) excised chromosome /al b sp</i> | white eye | 264 |
| re-insertion | <i>b Mhc[F02056], KG07247(w+), γ-ray 6 sp / al b sp</i> | red eye, black, speck | 20 |
| male recombinant | <i>KG07247(w+), γ-ray 6 sp / al b sp</i> | red eye, speck | 2 |
| total: | | | 3120 |

Table 4.2

Table 4.2: Different genotypes recovered from the Male recombination protocol implemented to remove the *MhcF02056* allele from the *Df pncr003;2L* genetic background. The excision and re-insertion classes correspond to natural transposition events of the P-element mediated by the $\Delta 2$ -3 transposase.

3- Discussion

In this chapter I have described the use of a genetic tool, the transposable element, to generate a null condition for the *pncr003;2L* smORF gene whose function I intended to characterise. However, this work took an unexpected detour, with the identification of a morphological phenotype associated with the line carrying the transposable element in question, but caused by a *Mhc* allele (*Mhc*^{F02056}) associated with the insertion bearing chromosome. In the context of the general usage of transposable elements as mutagenic reagents, this work exemplifies the issues that can arise from these kinds of background mutations. In this case the effect of the associated allele was particularly damaging, and unfortunate, because the phenotype it generated fitted well with the expression pattern and putative function of the *pncr003;2L* gene. This is one of the reasons why the characterisation of that background allele had to be so thorough, with that particular purpose having required a large amount of time, and the largest part of this chapter. The other reason is that the *pBac*{WH} *F02056* line was essential for the mutagenesis methods on *pncr003;2L*, being the only available line that could be used to disrupt that particular smORF gene, therefore the effects of this insertion, and the chromosome carrying it, had to be fully understood.

Regarding the *Mhc*^{F02056} allele, the main purpose of this work was to map the genetic locus of this initially unmapped allele, in order to understand its relationship with *pncr003;2L*. In this respect, this work was successful, as the different genetic and molecular mapping methods used here, show that this allele is associated with the *Mhc* gene. Beyond the genetic mapping of this allele this study suggests that this allele does not affect the protein sequence of the IFM-specific *Mhc* isoform, but instead may arise from a misregulation of *Mhc* alternative splicing. This misregulation is probably caused by a genomic deletion in the intronic sequence prior to the alternative exon 7c, as shown by these preliminary results.

The transposon-based mutagenesis strategies used here, were effective, producing the required null condition for *pncr003;2L*, which can now be used for the functional characterisation of *pncr003;2L*.

Here I showed that this null condition has no apparent behavioural phenotype, since the *pncr003.2L* null flies have a normal motility in general, and their ability to fly seems to be unaffected. The ultra-structure of the myofibrils in indirect flight muscles, which have been shown to express these genes, also seem normal, showing that the *Df pncr003;2L* mutants also lack a morphological phenotype. These results could be in accordance with the subcellular localisation of these peptides in the dyads, pointing to a specialised physiological role for this gene, which may affect muscle function in a more subtle way, and would therefore require a more sensitive and detailed characterisation.

Chapter V- Using the *Drosophila* adult heart as a system for the phenotypical characterisation of *pncr003;2L*.

1- Introduction:

Through the work presented so far in this thesis I have shown that *pncr003;2L* codes for two peptides, which are expressed in somatic and cardiac muscles, where they localise to the dyads, a structure closely linked with the regulation of muscle contraction, and I have therefore proposed that these peptides may have a physiological function in the regulation of that process. On the other hand I have described the successful generation of a null condition for *pncr003;2L*, and have shown that this mutant does not present any morphological, nor gross behavioural phenotype, focusing on the indirect flight muscles, and the flight capabilities of the adult flies, respectively. These results have led me to suggest that the *pncr003;2L* gene may have a function in muscle contraction which may be too subtle to be detected by the gross, morphological and behavioural assay presented so far. Therefore, in order to characterise the function of this gene, there clearly seems to be a need to perform a more sensitive muscle function assay, which could identify subtle defects in muscle contraction.

An aspect which has not yet been explored, is the shift of expression of the *pncr003;2L* gene, which in adult abdomens, ceases to be expressed in somatic muscles, but remains strongly expressed in the dorsal cardiac vessel, indicating that this gene may have a function in the contraction of the adult heart. This is particularly interesting in relation to the need of a more specific muscle contraction assay, because the adult *Drosophila* heart represents an accessible system in which to study muscle contraction. The *Drosophila* heart has long been considered as a powerful genetic system in which to

study the development of this organ, particularly since the identification of the *tinman* (*tin*) cardiogenic gene in *Drosophila* [74]. The identification of *tin* in *Drosophila* led to the cloning of the *Nkx2-5* gene in vertebrates [75], representing one of the first steps in the recognition of the highly conserved developmental pathway of the cardiac system between fruit flies and humans [76]. Most recently, the development of functional assays, which allow the monitoring of cardiac activity in adult flies [26,27,28,29], has given rise to the emergence of *Drosophila* as an exceptional tool in which to study cardiac function and disease. The methods developed by the Bodmer lab, which allow to measure specific parameters such as heart rate, contractility and rhythmicity in dissected adult hearts, are particularly appealing as they do not require any specialised electrophysiological equipment to be implemented, while being sensitive enough to detect even heart defects due to the background genetic variation in *Drosophila* [28]. A study carried out by the Bodmer lab, on the *KCNQ* potassium channel in *Drosophila* [27], is particularly interesting with respect to my own work. In that study it was shown that mutations in *KCNQ*, the *Drosophila* homologue of human *KCNQ1*, which is involved in myocardial repolarisation, and is associated with Torsades des Pointes arrhythmias and sudden death [77], also causes heart arrhythmias in *Drosophila*. That result is particularly interesting because it shows that the core mechanisms leading to the human disease are conserved in flies. On the other hand, regarding *pncr003;2L*, it is also interesting that those mutant flies with severe heart arrhythmias do not display any other morphological nor behavioural phenotypes. This is significant because it shows that the heart is a system that may be sensitive enough to detect a phenotype for *pncr003;2L*.

Because of the subcellular localisation of the *pncr003;2L* peptides to the dyads, which is where the Ca^{2+} exchange that triggers muscle contraction and relaxation occurs, it is possible, if not likely, that these peptides may have a function in the regulation of Ca^{2+} cycling during muscle contraction. The adult *Drosophila* heart would also offer the possibility to test this hypothesis because, in combination with the genetically encoded GCaMP Ca^{2+} reporters, the adult heart has been shown to represent a good system in which to measure calcium handling, as shown by Lin *et al.* [34]. In their study, Lin *et al.* used a version of the GCaMP Ca^{2+} reporter to detect subtle Ca^{2+} transient differences in mutants for the Ca^{2+} responding TpnI protein.

In this chapter, I take advantage of the *Drosophila* adult heart system to demonstrate that *pncr003;2L* has a function in cardiac contraction. I show that even though *Dfpncr003;2L* mutants show no morphological or structural defects in cardiac muscles, they present significantly more arrhythmic heart contractions than their wild-type counterparts. I demonstrate that this phenotype is specific to *pncr003;2L*, by means of genetic rescues, which also allow to demonstrate that both peptides encoded by *pncr003;2L* have a function which is equivalent. Furthermore, using the GCaMP3 Ca^{2+} reporter, I show that *pncr003;2L* has a function in the regulation of Ca^{2+} cycling during muscle contraction in cardiomyocytes.

2- Results:

2.1- *pncr003;2L* null flies do not display any morphological abnormalities in the adult heart.

It has already been shown in this work that *pncr003;2L* does not affect the morphology of IFMs (Chapter IV, Figure 4.14), however that result cannot be extrapolated to all muscles, therefore the effects of *pncr003;2L* remain to be assessed in other kinds of *pncr003;2L* expressing muscles, including cardiac muscles. For such assessment a similar morphological analysis to that performed with IFMs ([Chapter IV](#)) was carried out focusing on cardiac muscles. The dorsal vessel of the adult fly is composed of an unchambered thoracocephalic aorta and an abdominal, multi-chambered contractile heart marked by incurring sets of alary muscles, which have a suspensory function [78], and ostia cells, which serve as valves to allow the exchange of hemolymph from the abdominal cavity into the heart [79] (Figure 5.1A). The contractile abdominal heart vessel itself is composed of a monolayer of myoepithelial cells, or cardiomyocytes, and is covered ventrally by a layer of longitudinal, non-cardiogenic muscles, which also have a suspensory function. Epifluorescence microscopy imaging of dorsal preparations of adult abdomens, using phalloidin-rhodamine to label all muscles, was used to assess the general structure of the cardiac tube and surrounding somatic muscles (Figure 5.1 B and C). At this level the *pncr003;2L* null flies do not display any visible defects, with the hearts and surrounding somatic muscles of both wild-type and mutant flies having similar sizes and overall appearance. Confocal microscopy imaging, which allows to observe the organisation of sarcomeres with higher resolution, also shows that the ventral longitudinal muscles (Figure 5.1 D and E), and the cardiomyocytes themselves

(Figure 5.1 F and G), have a sarcomeric organisation, which is on the whole comparable to that of wild-type hearts. This apparently wild-type sarcomeric organisation was also assessed at the ultra structural level using TEM (Figure 5.1 H and I). The TEM micrographs, taken with a 5000X magnification, show that the Z discs and M bands display similar organisation (distance between them) and densities, in the sarcomeres of cardiac myocytes of both, wild-type and *pncr003;2L* null hearts. These observations indicate that the absence of *pncr003;2L* does not affect the structure of the cardiac myofibrils, which is in agreement with this gene not having an effect in the overall morphology of the heart, and is in line with the previous results presented in this work, showing that *pncr003;2L* has no apparent structural functions in IFMs.

Figure 5.1: The *Df pncr003;2L* null flies show no structural or morphological defects in heart muscles.

(A) Diagram representing the dorsal cardiac vessel in the abdomen of adult flies, which is composed of an unchambered thoracocephalic aorta, and an abdominal multi-chambered contractile heart marked by incurring ostia and sets of alary muscles. This diagram was modified from [78] . (B-C) Epifluorescence images of phalloidin-rhodamine stained dorsal abdominal segment, showing that the heart of *Df pncr003;2L* null flies (C) is morphologically similar to wild type flies (B). (D-G) Confocal Fluorescence micrographs of adult heart structures stained with phalloidin-rhodamine, showing that the sarcomeric organisation of (D, E) longitudinal ventral muscles, and (F-G) cardiomyocytes is similar between (D, F) wild type hearts and (E, G) *Df pncr003;2L* mutant flies. Scale bars: 5µm. (H, I) TEM micrograph of sarcomeres from cardiomyocytes of (H) *Or-R*, or (I) *Df pncr003;2L* flies, showing that the ultrastructure of the cardiomyocyte myofibrils is comparable between wild type and *pncr003;2L* null flies. Mitochondria (m), A bands, I bands, and Z discs are indicated. Scale bar: 0.5µm.

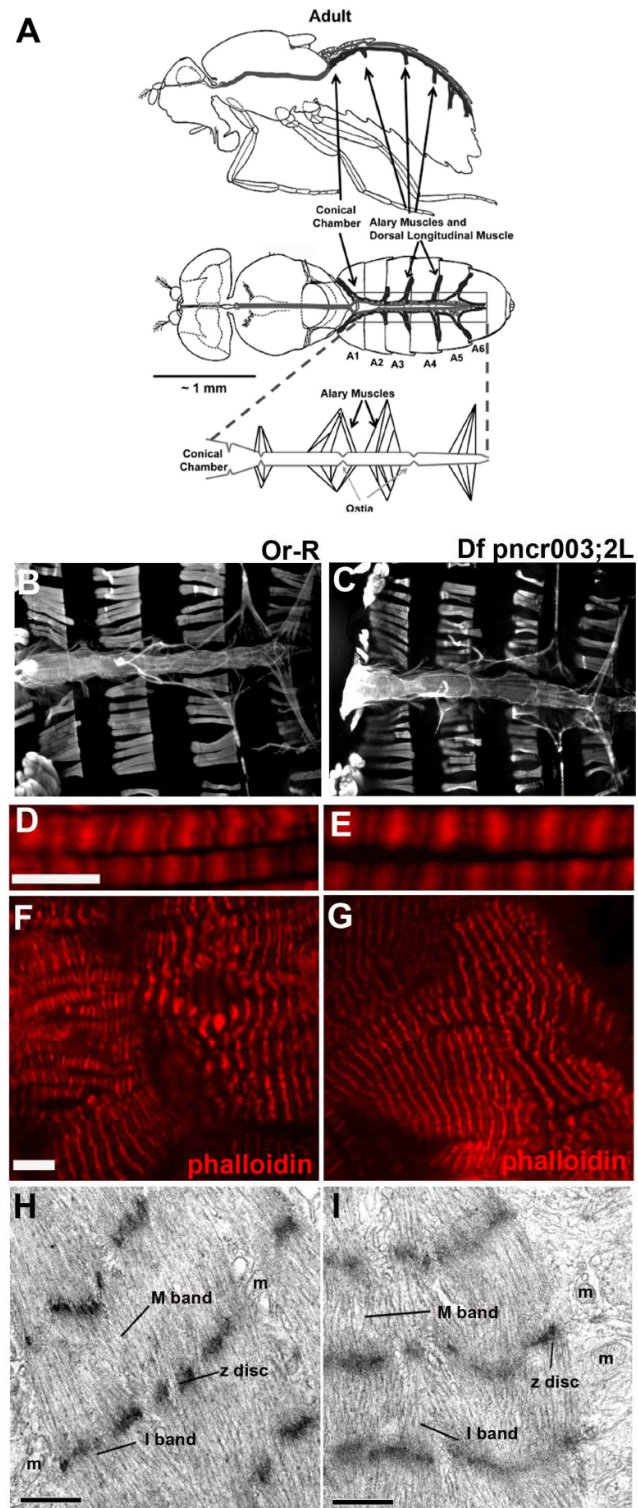


Figure 5.1

2.2- Recordings of live beating hearts show that *Df pncr003;2L* flies display arrhythmic heart contractions.

Since the morphology of *pncr003;2L* expressing muscles, whether IFMs or cardiac muscles, seems normal, and since the peptides encoded in this smORF gene localise to the dyads, I have hypothesised that they may have a physiological function. In this case, the phenotype of these mutants may become apparent if one focuses on how the muscles function, rather than on how they look. As a first approach to assess muscle contraction, I used a method previously described in the literature, used to record the endogenous contraction patterns of living heart preparations, which can essentially provide an indication of cardiac muscle function [26,27,29]. The power of this method is that because such patterns, in these preparations, are generated by the endogenous activity of the heart, through pacemaker cells located at the caudal end of the heart, and in the most anterior of the heart chambers, known as the conical chamber [80], these patterns can be recorded without the need to set up an electrophysiology setting to stimulate the muscles. The method consists of dissecting the flies in order to expose the contractile cardiac tube, while these are bathed in oxygenated *Drosophila* hemolymph saline [29]. In these conditions, it has been reported that the heart can remain beating for several hours after dissection —provided the saline solution is regularly replaced with fresh one [27,29]. In order to obtain this preparation, the abdomen of the fly is isolated from the rest of the body, thereby removing all components of the central nervous system. The ventral abdominal cuticle, gut and fat body are then carefully excised in order to expose the cardiac tube. The removal of the central nervous system is important, because unlike the myogenic, innervation-free larval heart, the adult heart is not entirely myogenic; it is innervated by Glutamatergic nerve terminals, which have been shown to be responsible for a phenomenon known as cardiac reversal, by which the peristaltic heart contractions which occur normally from the posterior to anterior ends is reversed [81]. By removal of CNS input these semi-intact preparations therefore allow us to study the intrinsic contractile activity of the heart [27]. After dissection to expose the heart, video recordings are then taken from these semi-intact preparations and used to produce a time-space-plot, also known as kymograph, which gives an account of the heart contraction patterns over an established period of time (of 20 seconds, for all the recordings presented in this work) (Figure 5.2 A and B).

Interestingly, the kymographs from wild-type and *Df pncr003;2L* mutant flies show a very striking difference: The latter appear to present irregularities in the period lengths of their heart contractions, contrasting with the regular contraction patterns of wild-type flies (Figure 5.2 A and B). Arrhythmic heart contractions in *Drosophila* have previously been described using a similar method, for mutants of the gene coding for the potassium channel alpha subunit *KCNQ1* homologue [27]. In that case, the arrhythmic behaviour was quantified using a parameter called “arrhythmicity index”, which divides the standard deviation of period lengths, by the period length median. Using this same metrics, a significant difference was observed between wild-type and *pncr003;2L* null flies, which present an arrhythmicity index two-fold that of wild-type hearts (Figure 5.2 D). These arrhythmic events do not seem to affect the overall heart frequency (Figure 5.2C), which remains very similar between wild-type and mutant hearts. Another parameter that can be obtained from these video recordings is the fractional shortening of the heart, which assesses the contractility capabilities of the cardiomyocytes by taking into account the difference in diameter between the most relaxed state (diastolic diameter) and the most contracted state (systolic diameter) of the heart. The hearts of *pncr003;2L* null mutants and wild-type flies also showed no significant differences in fractional shortening, indicating that *pncr003;2L* does not have a major effect on the contractility of cardiac muscles (Figure 5.2C). Importantly, the increased arrhythmicity was not observed with either of the two deficiencies that give rise to the *Df pncr003;2L* when these were assessed as heterozygotes (*Df γ-ray 6* /+, and *Df(2L)12* /+). These results show that the observed arrhythmicity is specifically induced by the synthetic homozygous deficiency *Df pncr003;2L*, which completely removes *pncr003;2L* (Figure 5.2 D). Because of the previously described allele affecting *Mhc*^{F02056}, associated with the deficiency *Df γ-ray 6*, it was necessary to rule out that this arrhythmic phenotype may be influenced by such allele. It is important to point out, however, that such an effect would seem unlikely because that particular *Mhc* allele is clearly recessive and seems to be specific to indirect flight muscles ([Chapter IV](#)).

Figure 5.2: The hearts of *Df pncr003;2L* mutants present heart arrhythmias. (A) Diagram representing the dorsal cardiac vessel in the abdomen of adult flies, showing the abdominal segment where measurements were taken, and still frames of a video recording of wild-type hearts with the heart in a diastolic, relaxed state, (left) and in a systolic (contracted) state. (B) Kymographs showing the pattern of heart contractions for wild-type and *Df pncr003;2L* mutant hearts. *Pncr003;2L* null hearts show irregular periods, some being abnormally long(asterisk). A normal heart period is indicated (green). (C) A quantification of heart frequency and fractional shortening show no significant difference between wild-type and *pncr003;2L* null flies regarding these parameters. (D) A quantification of the arrhythmicity index (standard deviation of heart period / heart period median) shows that *pncr003;2L* null flies present a significantly higher arrhythmicity than wild-type flies, as determined by a two-tailed Mann Whitney statistical test, (U=202, p<0.005). No significant difference was detected in either of the two deficiencies (*Df(2L)12*, and *Df γ-6*) that generate *Df pncr003;2L*, when tested as heterozygous. n=15-20 flies per genotype.

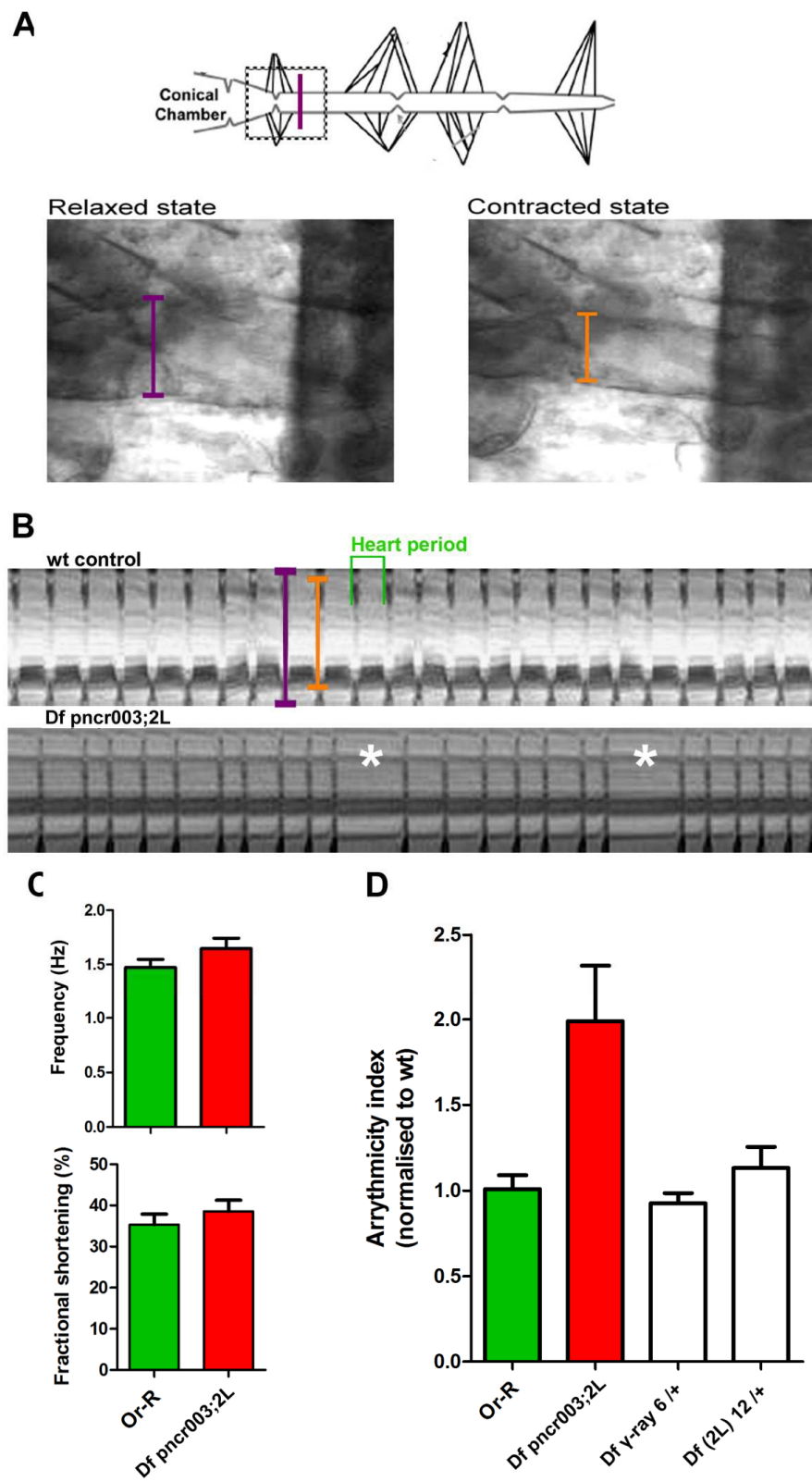


Figure 5.2

In order to determine whether the arrhythmic phenotype is independent from the heterozygous *Mhc* allele, the same arrhythmicity analysis was repeated with homozygous *Df(2L)12* flies carrying the genomic rescue for *CG31739* (*Df(2L)12,gCG31739*), and a *Df pncr003;2L* fly, which carries the male recombinant *Df* γ -ray 6 chromosome free from the *Mhc*^{F02056} allele (*CG17928*^{KG07247}, *Df* γ -ray 6 / *Df(2L)12*). In both cases, very similar arrhythmicity indexes were observed as for the original *Df pncr002;2L*, which confirms that the *Mhc*^{F02056} allele is not involved in this phenotype (Figure 5.3 A and B). Furthermore, when an RNA interference construct specific to *pncr003;2L* is expressed in muscles with the *Dmef2-Gal4* driver, which effectively reduces the expression of *pncr003;2L* (Figure 5.3 C), a very similar level of arrhythmicity is observed (Figure 5.3 A and B). This result shows that the loss of function of *pncr003;2L* itself seems to be at the origin of this arrhythmicity, which is important, considering that all of the above mentioned genomic deletions also remove two other genes (*CG13282*, and *CG13283*) which could also be responsible for the phenotype.

2.3- *pncr003;2L* rescues the arrhythmicity phenotype

If the arrhythmic cardiac contractions presented by the *Df pncr003;2L* synthetic homozygous deletion and the different genetic conditions described above are exclusively due to the loss of function of *pncr003;2L*, it would be expected that this particular phenotype would be corrected by restoring the expression of *pncr003;2L* in the *Df pncr003;2L* mutant background.

To test this, I generated a series of constructs designed to induce the expression of different versions of *pncr003;2L*; all under the control of the UAS promoter, and all used to generate transgenic flies with the PhiC31 integrase mediated system [40], with which all transgenic constructs are inserted in the same specific site within the genome. This site specific transgenesis method ensures that the expression of these different constructs is not influenced by any positional effects [40], therefore any functional difference between them can be attributed to the nature of the construct itself.

One of these constructs, called simply *pncr003;2L*, expresses the AB isoform, carrying both ORFs A and B (see [Chapter III](#)). This construct was obtained by cloning the RE28911 cDNA into the *pUASattB* vector, and represents the full length version of one of the *pncr003;2L* transcripts, as endogenously expressed in the heart and somatic muscles. Another construct, carries a version of the same AB isoform in which a

frameshift has been introduced in both ORFs A and B (*pncr003;2L FS*). This transcript was obtained by *de novo* custom sequence synthesis (by Eurogentec), with the aim of using it as a control, in order to distinguish whether the function of the *pncr003;2L* gene is carried out by its two encoded peptides, whose amino acid sequences have been completely lost because of the frameshift, or by the RNA transcript itself, whose sequence remains relatively intact since only a few point mutations were required to create the frameshifts (Annex 3). When expressed in the muscles of *Df pncr003;2L* mutant flies using the *Dmef2-GaL4* driver, the *pncr003;2L* construct rescued the arrhythmicity to wild-type levels. On the other hand, the *pncr003;2L FS* construct or the expression of the *Dmef2-GaL4* driver on its own had no effect on the mutant arrhythmicity phenotype (Figure 5.4A and B).

These results show that the arrhythmicity is due to the loss of *pncr003;2L* and, more specifically, due to the loss of its peptide sequences. Using a heart specific driver (*tinman-GaL4*), which is expressed exclusively in the contractile cardiomyocytes [82,83] instead of the pan-muscular driver *Dmef2-GaL4*, also rescues the phenotype when used to drive the *pncr003;2L* construct, but fails to do so with the *pncr003;2L FS* construct (Figure 5.4B). This indicates that the function of the *pncr003;2L* is required specifically by cardiac cells, which I have shown to express the *pncr003;2L* transcripts (Chapter III). Although the *pncr003;2L AB* transcript carries both ORF A and ORF B, it was shown in Chapter III, that this transcript does not seem capable of polycistronic translation *in vivo*. This suggests that, in this case, the *pncr003;2L* ORF A peptide on its own, would be responsible for the observed rescue, and that this one peptide would therefore be sufficient to convey the function of the *pncr003;2L* gene in this context.

In order to test this and also to assess functionally each peptide, two more constructs carrying only one of the two ORF sequences each, were generated. These two constructs, called *pncr003;2L ORF A*, and *pncr003;2L ORF B*, only carry the fragment of the transcript corresponding to the ORF sequence including a stretch of some 100 nt upstream of each of them, which includes their endogenous translation context. Interestingly, either of these two constructs was sufficient to rescue the arrhythmic phenotype (Figure 5.4B), indicating that, at least in this specific context, both ORFs convey a very similar function. Furthermore, each of the N-terminal FLAG-Hemagglutinin tagged peptides constructs (FH-ORF A and FH-ORF-B), also rescue this phenotype (Figure 5.4B), which not only confirms that both peptides are functionally

equivalent, but also proves that the N-terminal FLAG-Hemagglutinin tag does not affect the function of the peptides and therefore that the subcellular localisation, in the dyads of the muscle cells, reported by these constructs in [Chapter III](#) is probably genuine.

Since the results obtained so far indicate that the loss of function of *pncr003;2L* results in heart arrhythmicity, I assessed whether the excess of function of these peptides would have any effect. Surprisingly, when expressed in a wild-type background both peptides, ORF A and ORF B, also give rise to an arrhythmic behaviour, similar to that observed with *pncr003;2L* loss of function (Figure 5.5A and C). In this case, similarly as with the rescue experiments, no significant increase in arrhythmicity is detected when a frameshift-carrying construct is used in the same conditions of ectopic expression (Figure 5.5A and C). The other parameters assessed for the *pncr003;2L* loss of function, such as frequency and fractional shortening, also seem unaffected by over-expression of either peptide, with respect to the *pncr003;2L FS* control line (Figure 5.5B). These results indicate that either loss or gain of function of the *pncr003;2L* peptides result in heart arrhythmias, suggesting that the process that involves these peptides is sensitive to their dosage.

Figure 5.3: The arrhythmias presented by *Df pncr003;2L* are specific to *pncr003;2L* and independent from the *Mhc^{F02056}* allele. (A) Kymographs showing the pattern of heart contractions in wild-type flies (Or-R), flies expressing a *pncr003;2L RNAi* construct in muscles, using *Dmef2-GaL4* as a driver, and in the two null conditions for *pncr003;2L* free of the *Mhc^{F02056}* allele (Df(2L)12, gCG31739 and Df(2L)12 / KG7247, Df γ -6). Similar irregular heart periods, as those observed in *Dfpncr003;2L*, are displayed by the RNAi expressing line, and by the two null conditions for *pncr003;2L* free of the *Mhc^{F02056}* allele, showing that this phenotype is specific to the removal of the *pncr003;2L* locus. (B) A quantification of the arrhythmicity index between these genotypes shows that the *pncr003;2L RNAi* knock-down and the (Df(2L)12, gCG31739 and Df(2L)12 / KG7247, Df γ -6) *pncr003;2L* null conditions present a significantly higher arrhythmicity than wild-type flies, as determined by a two-tailed Mann Whitney test, (U=57, p<0.005,), (U=46, p<0.05) , and (U=9, p<0.002). n=10-15 flies per genotype. (C) Semi-quantitative RT-PCR on mRNA extracts of whole flies, showing that the expression of *pncr003;2L* is visibly reduced by the expression of the *pncr003;2L RNAi* construct (using the same primers as in Figure 4.2C).

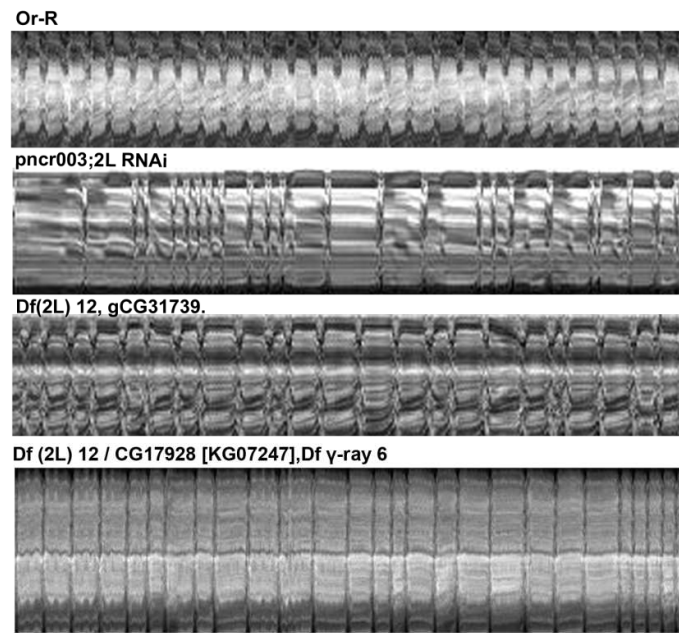
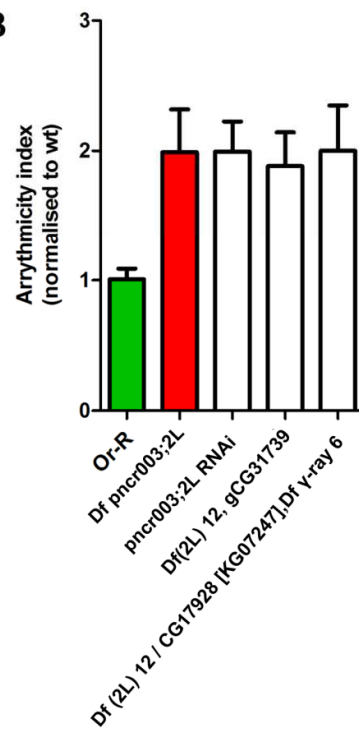
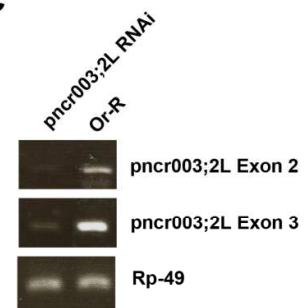
A**B****C**

Figure 5.3

Figure 5.4: The arrhythmias presented by *Df pncr003;2L* are corrected with different *pncr003;2L* expression constructs. (A) Kymographs showing the pattern of heart contractions in *Df pncr003;2L* flies expressing, in muscles, either a *pncr003;2L* rescue construct (*Df pncr003;2L, Dmef2>pncr003;2L*), or the frame shift carrying *pncr003;2L-FS* control (*Df pncr003;2L, Dmef2>pncr003;2L FS*). The heart arrhythmicity phenotype observed in *Df pncr003;2L* mutants is corrected in flies expressing the *pncr003;2L* rescue construct, but not in flies expressing the *pncr003;2L-FS* control. (B) A quantification of the arrhythmicity index shows that the flies expressing the *pncr003;2L* rescue construct have a significantly lower arrhythmicity index than *Df pncr003;2L* mutants ($U=160$, $p<0.05$), whereas flies expressing the *pncr003;2L FS* control construct, or carrying the *Dmef2-GaL4* driver but no UAS-expression construct have no significant effect. The arrhythmicity is also significantly corrected in *Df pncr003;2L* flies expressing constructs carrying only the ORFA, or ORF B sequences, tagged with the N-terminal FLAG-Hemagglutinin tag (*Df pncr003;2L, Dmef2>FH_ORFA*, and *Df pncr003;2L, Dmef2>FH_ORFB*), or not ((*Df pncr003;2L, Dmef2> ORFA*, and *Df pncr003;2L, Dmef2> ORFB*), ($U=103$, $p<0.05$), ($U=97$, $p<0.05$), ($U=61$, $p<0.05$), ($U=110$, $p<0.005$), respectively. (C) The arrhythmicity is significantly corrected by using the *tinman-GaL4* cardiac specific driver instead of the pan-muscle driver *Dmef2-GaL4*, to express the *pncr003;2L* rescue construct (*Df pncr003;2L, tin>pncr003;2L*), as determined by a two-tailed Mann Whitney statistical test, (($U=105$, $p<0.0005$), but not the the *pncr003;2L FS* control construct (*Df pncr003;2L, tin>pncr003;2L*). $n=15-20$ flies per genotype.

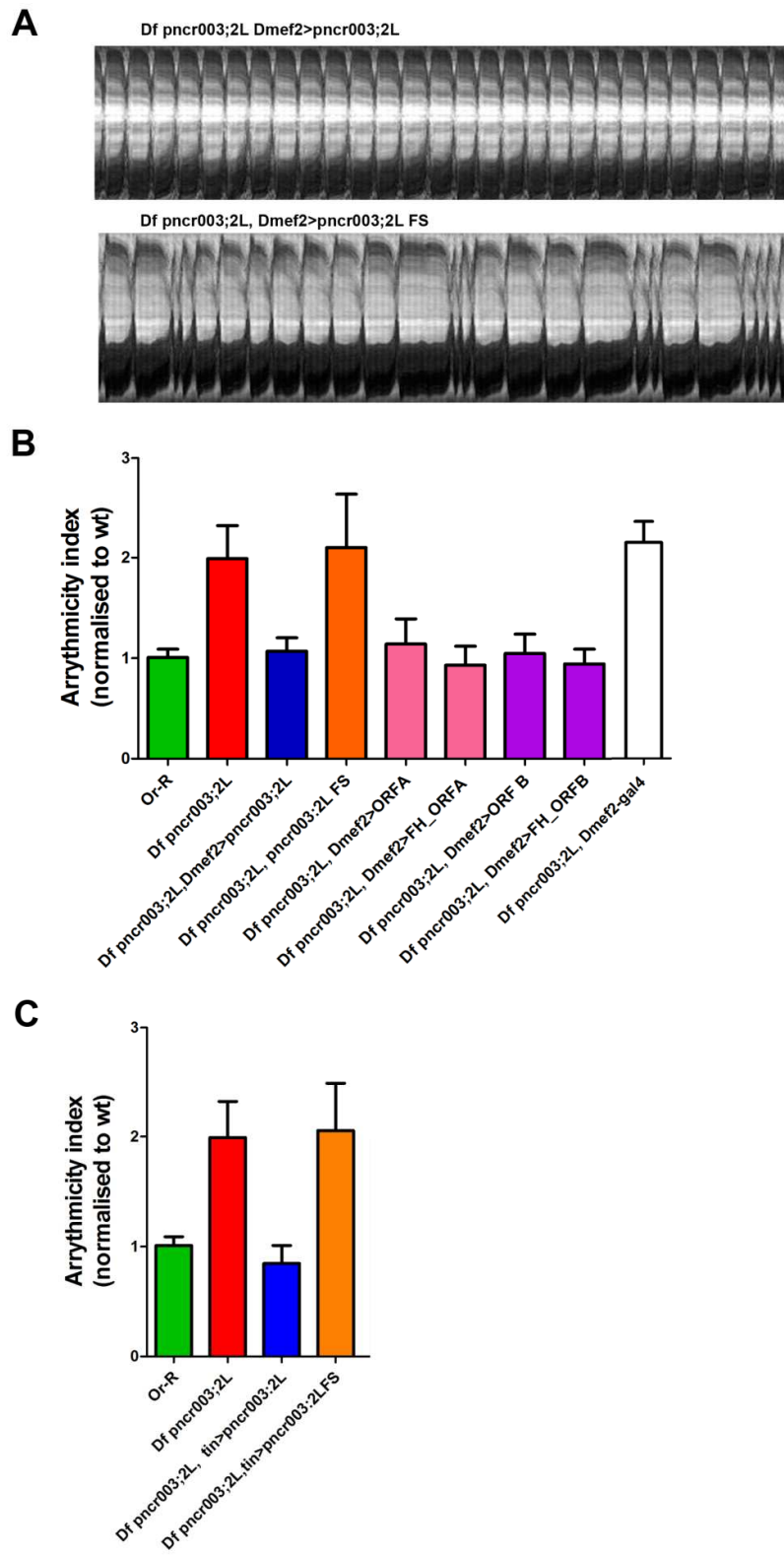


Figure 5.4

Figure 5.5: *pncr003;2L* excess of function also leads to heart arrhythmia. (A)

Kymographs showing the pattern of heart contractions in flies expressing, in muscles, either the *pncr003;2L FS* control construct (*Dmef2>pncr003;2L FS*), or the constructs expressing either *pncr003;2L ORFA* (*Dmef2> pncr003;2L ORFA*) or *pncr003;2L ORF B* (*Dmef2> pncr003;2L ORFB*). The over-expression of either the *pncr003;2L ORFA* or *pncr003;2L ORFB* constructs in a wild type background, leads to arrhythmic heart contractions, this is not the case when the construct is expressed in the same background. (B) A quantification *pncr003;2L FS* control of heart frequency and fractional shortening show no significant difference between the flies over-expressing the *pncr003;2L FS* controls and the flies expressing the *pncr003;2L ORFA* and *ORF B* transcripts, regarding these parameters. (C) A quantification of the arrhythmicity index shows that the flies expressing the *pncr003;2L ORFA* or *ORFB* constructs have a significantly higher arrhythmicity index than wild type flies, as determined by a two-tailed Mann Whitney statistical test, ($U=52$, $p<0.05$), ($U=108$, $p<0.05$), respectively, whereas flies expressing the *pncr003;2L FS* control construct have no significant effect. $n=15-20$ flies per genotype.

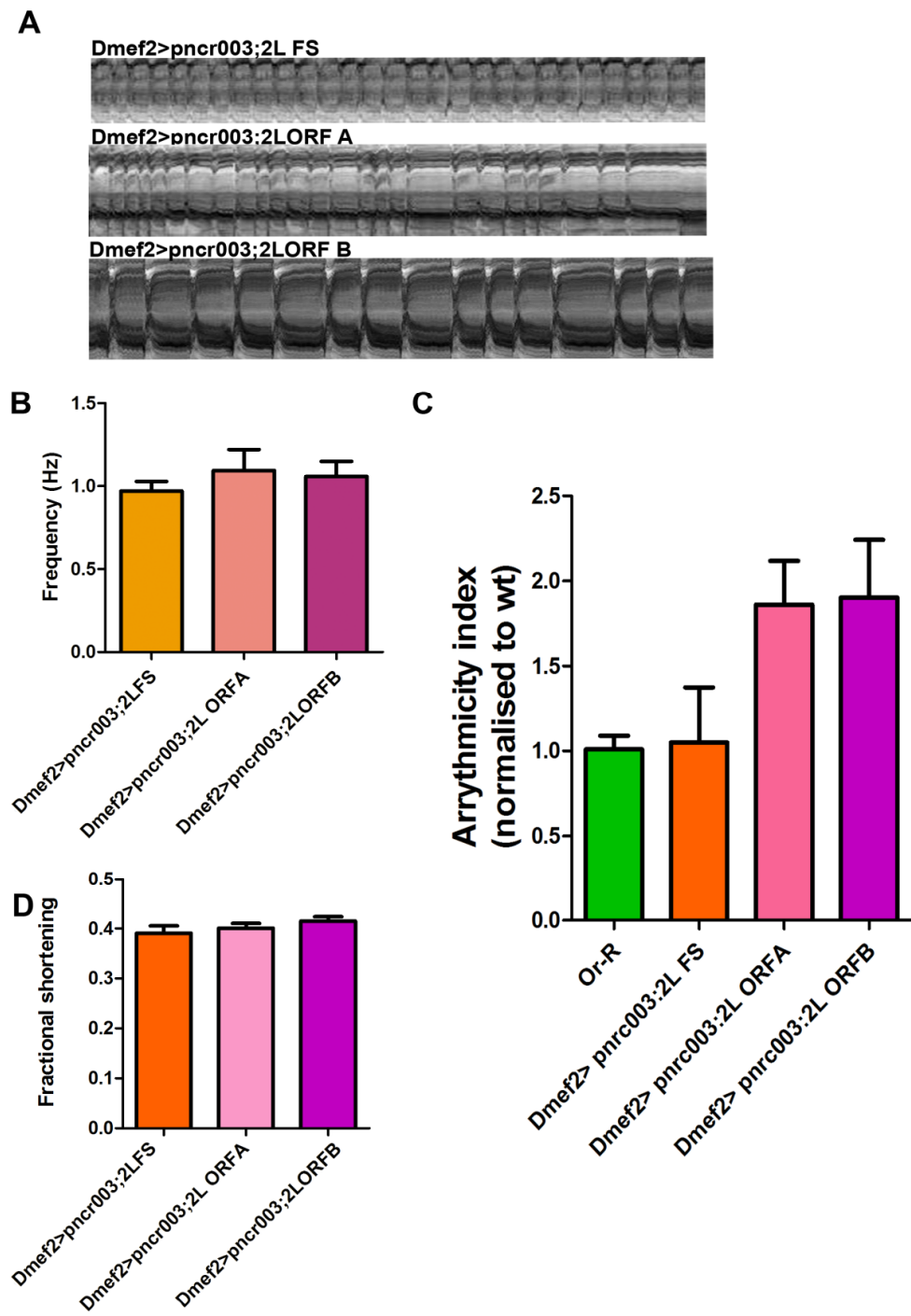


Figure 5.5

2.4- *pncr003;2L* mutants present abnormalities in their intracellular action potential recordings.

Data obtained by Jeremy Niven [33], who performed intracellular action potential (AP) recordings of cardiomyocytes in the *Df pncr003;2L* mutant lines in order to assess the physiological features of these cells, show that these mutants present specific abnormalities in their action potential patterns; while the wild type recordings show uniform patterns of APs, the mutant recordings sometimes show failed APs, and often show APs with a “double peak” appearance (Figure 5.6A). Interestingly, these abnormalities are not present in *Df pncr003;2L* animals expressing the *pncr003;2L* rescue construct, mirroring the results described above when assessing the cardiac arrhythmias. By quantifying the data provided by Jeremy Niven, I could show that the incidence of double action potentials is in fact significantly higher in *Df pncr003;2L* mutant hearts than wild type, or than in *pncr003;2L* rescued hearts (Figure 5.6B). The heart failure events do not have a significantly higher incidence in the mutants because of the high variability of these events (Figure 5.6C), which is reflected by a significantly different variance between mutant and wild-type animals regarding these particular AP failure events. Similar to what has been observed with the heart contraction video recordings, the mutant condition showed no differences in AP frequencies compared to wild type flies (Figure 5.6 D). No differences were observed in AP amplitudes either (Figure 5.6E). These results suggest that individual cardiomyocytes, with their abnormalities in AP patterns, reflect the arrhythmic behaviour observed when assessing the heart as a whole organ.

Figure 5.6: *pncr003;2L* mutants have abnormal action potential patterns. (A)

Sample traces of intracellular recordings, courtesy of J.E. Niven, from adult cardiomyocytes of wild-type (green); *Df pncr003;2L* (red); and *Df pncr003;2L* expressing the *pncr003:2L* rescue construct (*Df pncr003;2L, Dmef2>pncr003;2L*) (blue), showing that *Df pncr003;2L* flies have defects in their action potential patterns, which are corrected by the expression of *pncr003;2L*. Arrows indicate “double” action potentials (AP). Arrowheads indicate failed action potentials. Grey dashed line indicates resting potentials. A sample peak from each trace (underlined) appears magnified. (B-C) Quantification of action potentials from intracellular recordings, showing average percentage of action Potentials (APs) with the double peak phenotype (B) and the percentage of failed APs per cell (C). The percentage of double APs is significantly higher in *Df pncr003;2L* mutants, as assessed by a one-tailed Mann-Whitney U test, than in wild type ($U=14$, $p<0.005$), and than in *Df pncr003;2L* flies expressing the *pncr003;2L* rescue construct ($U=7$, $p<0.008$). Although the difference in number of failed APs between *Df pncr003;2L* and the other genotypes is not statistically significant according to a Mann-Whitney U test, due to the high variability of the phenotype; an F-test shows that the difference in variance is statistically significant ($F=20.42$, $p<0.0001$). Thus, failed APs are rare in wild-type but they do appear erratically in *Df(2L)scl* mutants. (D-E) No significant differences in neither frequency (D) nor action potential amplitude (E) could be observed between these genotypes. $n = 8-16$ cells per experiment.

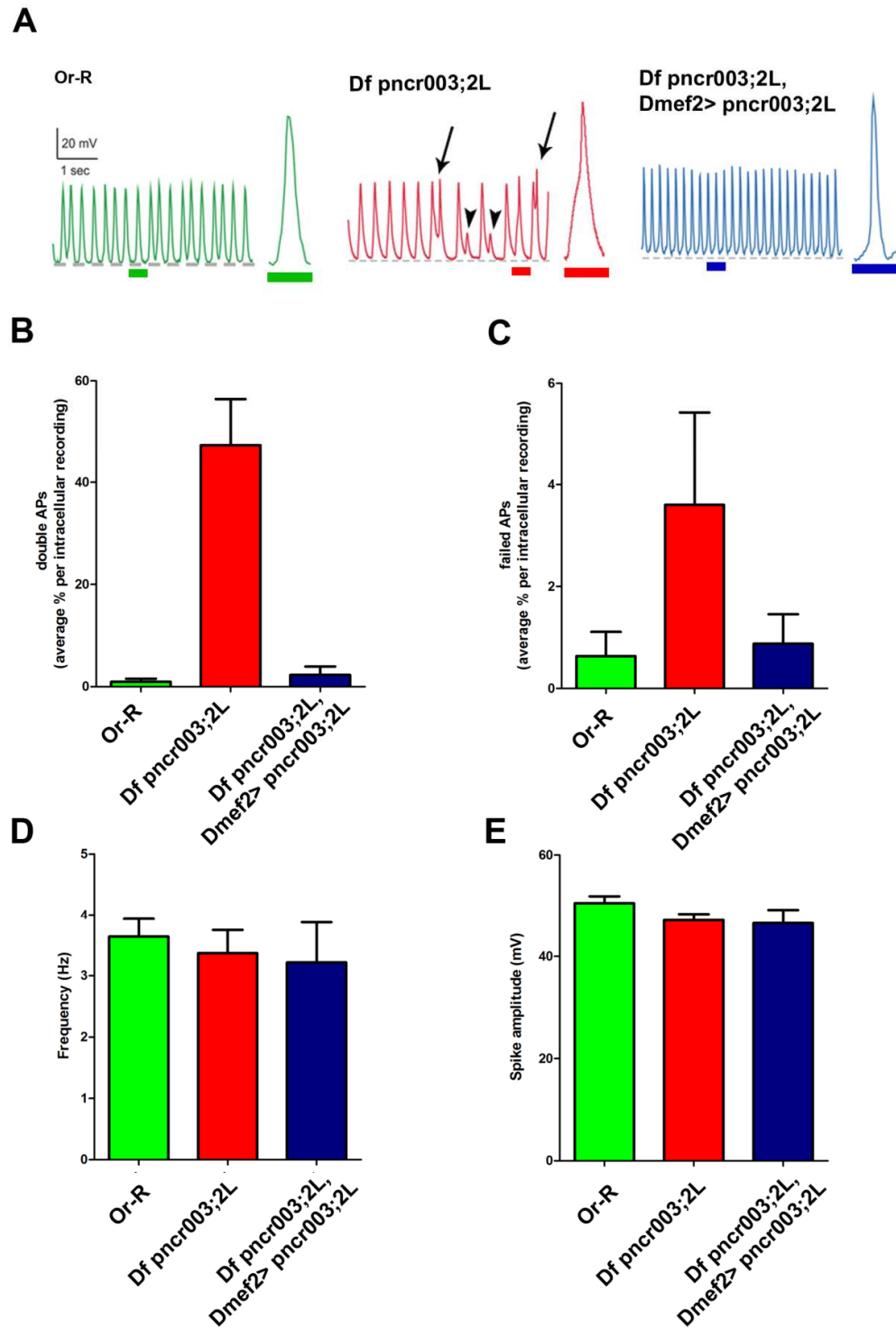


Figure 5.6

2.5- *pncr003;2L* influences calcium levels during heart contraction.

Having identified that *pncr003;2L* null flies have a defect in their heart contraction patterns, the next step in the functional characterisation of this gene would be to identify the underlying cause of this cardiac arrhythmia. In [Chapter III](#), I showed that the *pncr003;2L* peptides localise to the dyads in indirect flight muscles and cardiac muscles. As explained in that chapter, the dyads are a structure at the centre of the muscle contraction process. This structure regulates the release of Ca^{2+} from the sarcoendoplasmic reticulum into the cytosolic space, which is necessary to trigger the conformational changes upon its binding to Troponin C of the Troponin-Tropo-myosin complex. This binding leads to the exposure of myosin binding sites on the actin filament, necessary for the acto-myosin interaction resulting in muscle contraction. Given the localisation of the *pncr003;2L* peptides to that particular structure, one could hypothesise that these peptides may have a role in the regulation of Ca^{2+} during heart contraction. In order to test this possible involvement of the *pncr003;2L* peptides in the Ca^{2+} cycling process, I took advantage of the genetically encoded Ca^{2+} reporter G-CaMP3 [84], a chimeric fusion of the GFP and Ca^{2+} calmodulin (CaM) proteins, which acts as a fluorescent indicator of Ca^{2+} . In its Ca^{2+} free form, this chimeric protein has a conformation, which interferes with the chromophore domain of GFP, leading to a poor fluorescence signal, but in the presence of Ca^{2+} , the conformational change of the Ca^{2+} bound CaM restores the chromophore domain of GFP, leading to a significant increase in fluorescence signal [85]. I used the living heart preparations described above, in combination with the G-CaMP3 reporter, to compare the Ca^{2+} -dependent fluorescence signal, or Ca^{2+} transients, between wild-type and *pncr003;2L* null flies during heart contraction (Figure 5.7). This approach is similar to that used to measure Ca^{2+} handling during heart contraction by Lin *et al.* [34]. The main differences between the approach implemented here, and the one previously reported by Lin *et al.*, is that here confocal imaging instead of epifluorescence, is used to acquire time lapse series taken at 16.6 frames / second (fps), which although slower than the 100fps obtained by Lin *et al.*, is still sufficient to sample the contraction events occurring at 1-3 Hz. Also, and most importantly, here I use the G-CaMP3 Ca^{2+} reporter, which is significantly more sensitive, and yields less background noise than the GCaMP2 indicator used by Lin *et al.* [84].

The comparison between Ca^{2+} -dependent fluorescence recordings from wild-type and *Df pncr003;2L* hearts, in which GCaMP3 was expressed with the muscle specific *Dmef2-GaL4* driver, shows that the Ca^{2+} transients of *pncr003;2L* null hearts have a significantly wider amplitude, and steeper decay than wild-type (Figure 5.7A). These results indicate that *pncr003;2L* does indeed have a role in the regulation of calcium cycling during heart contraction. In order to plot average values of intensity over time for these transients, and because every heart has a unique beating frequency, for each peak, I focused on the decay phase, normalising the duration of the signal in relation to the maximum intensity value (representing the $t_{0\%}$ time point) and the lowest basal value (representing $t_{100\%}$). By focusing on the decay phase it was possible to fit a 2nd order polynomial curve to the data points of each peak, which were initially differentially distributed across the time axis, in order to obtain decay phase curves with the same data points for each peak, which can now be averaged (Figure 5.7 B). Following this procedure, *Df pncr003;2L* hearts show a calcium transient amplitude which is almost double that of wild-type hearts (Figure 5.7C).

A very similar difference in calcium amplitude, and decay can be observed between the Ca^{2+} transients of *Df pncr003;2L* hearts expressing the UAS-*pncr003;2L* rescue construct, and those of *Df pncr003;2L* hearts expressing the UAS-*pncr003;2L FS* control construct (Figure 5.8A), showing that the increment in amplitude of the Ca^{2+} dependent fluorescence signal, like the arrhythmic phenotype, depends on the expression of *pncr003;2L*. In line with these results, when the *pncr003;2L ORFA*, or the *pncr003;2L ORFB* constructs are over-expressed in a wild-type background, the calcium transients observed during heart contraction show the opposite effect to that observed with the *Df pncr003;2L* mutants. The calcium transients of *pncr003;2L ORFA*, or *pncr003;2L ORFB* expressing hearts show significantly decreased amplitude when compared with the transients of hearts expressing the *pncr003;2L FS* controls (Figure 5.8B). Overall, these results show that in agreement with their subcellular localisation in the dyads, the *pncr003;2L* peptides have a physiological role in the regulation of calcium cycling during the contraction of cardiac muscles, with the loss of function of these peptides increasing the amounts of calcium released upon contraction, and their excess of function reducing it. In these experiments, both the loss and excess of function of the *pncr003;2L* peptides, and therefore either higher or lower than normal amounts of calcium released upon muscle contraction, have been shown to produce arrhythmic

heart contractions, therefore, it could be hypothesised that the regulation of calcium needs to be tightly regulated in order to ensure the rhythmic contractions observed in wild-type hearts.

Figure 5.7: *pncr003;2L* null mutants present Ca²⁺ transients with higher amplitudes during heart contraction. (A) Sample fluorescence confocal images of GCaMP3 expressing hearts in systolic states, and their corresponding traces, showing raw calcium $\Delta F/F_0$ fluorescence intensities from 10 second recordings, for wild-type control hearts (*Dmef2>GCaMP3*) and *Df pncr003;2L* hearts (*Df pncr003;2L, Dmef2>GCaMP3*), showing that *pncr003;2L* null hearts have calcium transients with higher amplitudes. (B) Each peak was normalised in relation to its decay phase, normalising the duration of the signal in relation to the maximum intensity value (representing the $t_{0\%}$ time point) and the lowest basal value (representing $t_{100\%}$). By focusing on the decay phase it was possible to fit a 2nd order polynomial curve to the data points of each peak, which were initially differentially distributed across the time axis, in order to obtain decay phase curves with the same data points for each peak, which can now be averaged. (C) Averaged G-Camp3 fluorescence signals of calcium transients normalised as in (B), and plotted in relation to the wild-type average maximum signal. *Df pncr003;2L* hearts show significantly higher calcium transient amplitudes than wild type hearts, as determined by a two-tailed Mann Whitney statistical test, ($U=7$, $p<0.005$,**), $n=10$ flies per genotype.

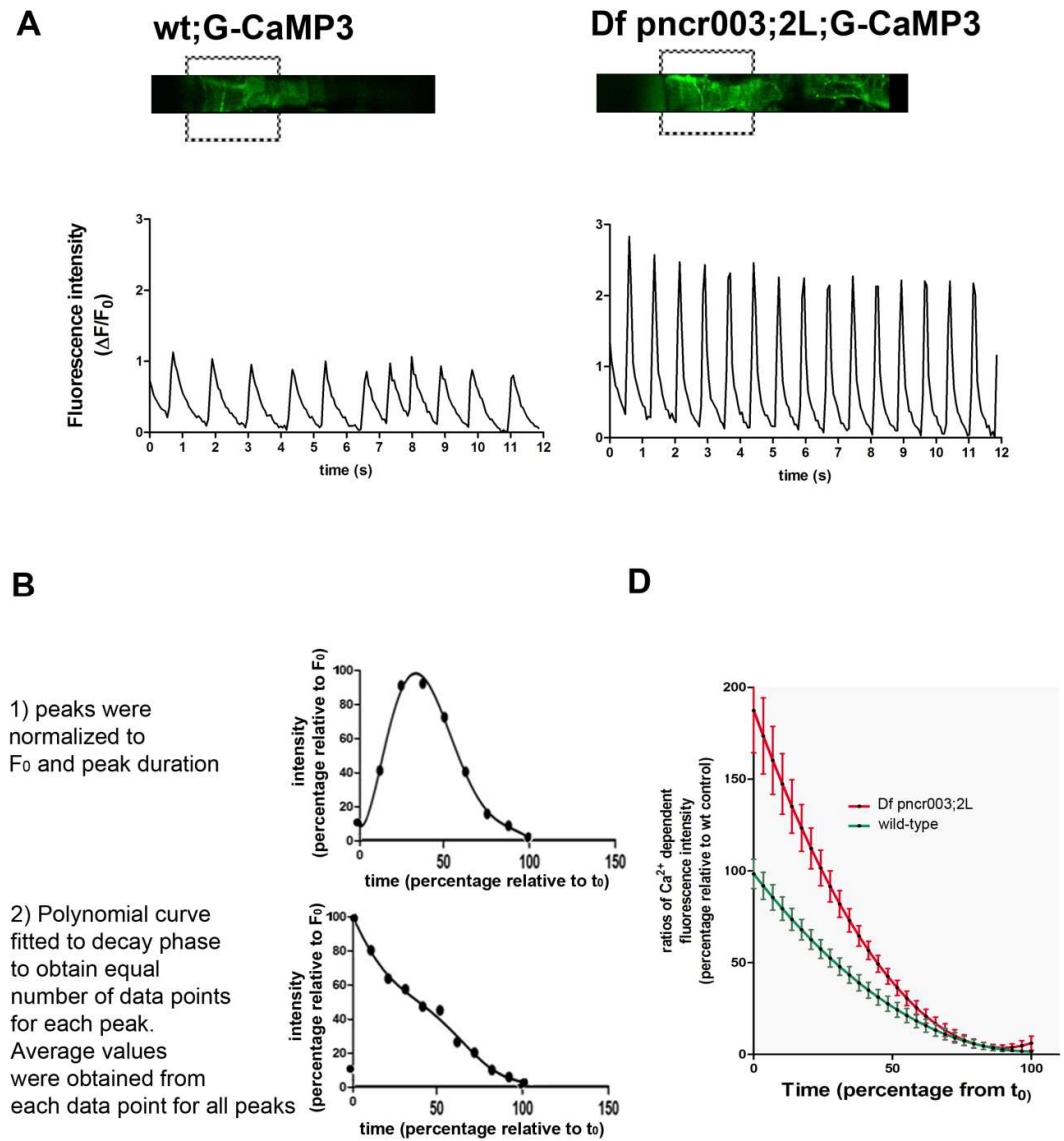


Figure 5.7

Figure 5.8: mutant rescues and over-expression effect of *pncr003;2L* in calcium

transients. (A-B) Averaged GCaMP3 fluorescence signals of Ca^{2+} transients normalised as in Figure 5.7B, and plotted in relation to the wild-type average maximum signal. (A) *Df pncr003;2L;GCaMP3* flies expressing the *pncr003;2L* rescue construct (*Df pncr003;2L, Dmef2 > pncr003;2L* hearts show significantly reduced calcium transient amplitudes, similar to those of wild type hearts, compared to *Df pncr003;2L;GCaMP3* mutants expressing the *pncr003;2L FS* control construct (*Df pncr003;2L, Dmef2 > pncr003;2L FS*), which have relative amplitudes comparable to those observed in *Df pncr003;2L* mutants. Statistical significance was determined by a two-tailed Mann Whitney statistical test, (U=5, $p < 0.05$), n=10 flies per genotype.

(B) over-expression of the *pncr003;2L ORFA* and *pncr003;2L ORFB* constructs in a GCaMP3 genetic background (*Dmef2 > pncr003;2LORFA* and *Dmef2 > pncr003;2LORFB*) show significantly reduced calcium transient amplitudes, compared to flies over-expressing the *pncr003;2L FS* control construct in the same genetic background (*Df pncr003;2L, Dmef2 > pncr003;2L FS*). Statistical significance was determined by a two-tailed Mann Whitney statistical test, (U=1, $p < 0.005$), (U=5, $p < 0.05$). n= 8 to 10 flies per genotype.

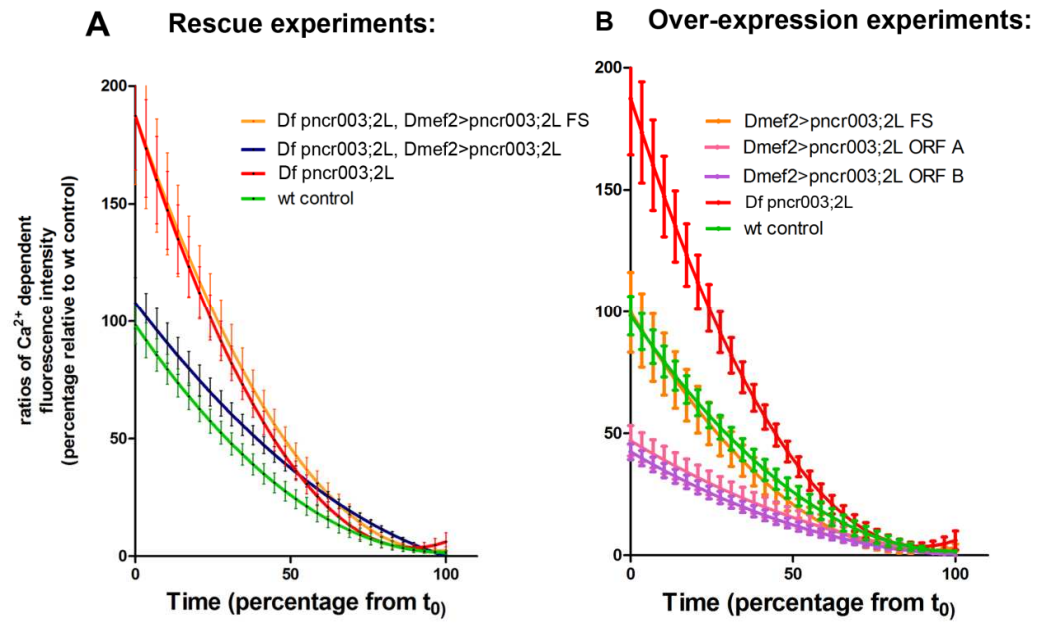


Figure 5.8

3. Discussion:

In this chapter, through the use of the adult heart as a system to study the effects of *pncr003;2L*, I provide experimental evidence showing that the two smORFs encoded in the *pncr003;2L* gene, have a function in the regulation of Ca^{2+} in cardiomyocytes. I have shown here that the *pncr003;2L* null mutants, generated in [Chapter IV](#), present heart arrhythmias, and calcium transients which have significantly higher amplitudes than wild type heart.

Throughout the work presented here, a strong emphasis has been placed on showing that these abnormalities, in heart rhythm and Ca^{2+} transient amplitudes are due to the lack of function of the *pncr003;2L* gene itself, and more specifically to the peptides it encodes.

First of all, I show that the heart morphology of *pncr003;2L* mutants, which was analysed at the whole organ, and at the ultra structural level, is comparable to that of wild type flies, indicating that the origin of the arrhythmias is physiological rather than morphological, which fits well with the phenomenology described throughout this thesis for the *pncr003;2L* gene, and most particularly with the localisation of its encoded peptides to the dyads. I show that the arrhythmia is also independent of the recessive *Mhc*^{F02056} allele described in the previous chapter of this thesis, and present, in an heterozygous condition, in the *Df pncr003;2L* null condition, since neither the *Mhc*^{F02056}-carrying deficiency *Df γ-6*, nor the deficiency *Df(2L)12*, used to generate the *Df pncr003;2L* null condition, produce the arrhythmicity phenotype as heterozygous. In line with these results, and showing that these phenotypes are not due to any sort of genetic interaction between a *pncr003;2L* null condition and the *Mhc*^{F02056} allele, the two *Mhc*^{F02056}-free, null conditions for *pncr003;2L* (the *CG17928*^{KG07247}, *Df γ-ray 6* and *CG31739 (Df(2L)12,gCG31739)*, as well as the strong reduction of the *pncr003;2L* expression by means of the RNAi knock-down, show the same arrhythmicity as *Df pncr003;2L*.

The major proof of specificity comes from the genetic rescues of the arrhythmicity and Ca^{2+} transient phenotypes by the different *pncr003;2L* constructs generated in this work. I have shown that the expression of the *pncr003;2L* construct, in either all muscles, with the *Dmef2-GaL4* driver, or specifically in hearts, with the *tin-GaL4* driver, is sufficient

to rescue the arrhythmicity phenotype. Showing that both, the heart arrhythmias, and the function of *pncr003;2L*, are linked to the intrinsic function of the *tinman* expressing cardiomyocytes, where the expression of *pncr003;2L* is detected.

This system, in which a phenotypic rescue by *pncr003;2L* is possible, allowed me to prove that the function of *pncr003;2L* is conveyed by its encoded peptides and not by the *pncr003;2L* mRNA itself, as shown by the lack of rescue—or over-expression phenotypes—when using the frame-shift carrying construct (*pncr003;2L FS*), in which only the smORF sequences are disrupted by a few punctual changes in the nucleotide sequence. Furthermore, this system allowed me to individually test the function of each of the ORFs encoded by the *pncr003;2L* gene. This was particularly interesting as the extensive similarity between these peptides appears to contrast with the tissue-specific expression patterns of some of the transcripts encoding them, as has been described in [Chapter III](#), and therefore the question of their functional equivalence remained open. My results favour the hypothesis that both of the peptides encoded by this gene have an equivalent function, at least in this specific context, since they were both able to rescue the mutant arrhythmicity and Ca^{2+} transient phenotypes, and to induce the same over-expression phenotypes. Importantly, I also show here that the FH-tagged peptides are able to rescue the *pncr003;2L* null phenotypes as well, proving that the tag does not interfere with their function, and therefore showing that the subcellular localisation they reflect is unlikely to be artifactual.

Regarding the effects generated by the lack, or excess of function of *pncr003;2L*, the fact that these two conditions have the opposite effect on the amplitude of the Ca^{2+} transients, and therefore in the Ca^{2+} dynamics during heart contraction, clearly points to a regulatory role of the *pncr003;2L* peptides over the Ca^{2+} cycling process. However, the relationship between the misregulation in calcium cycling, by either the lack or excess of function of *pncr003;2L*, and the observed arrhythmias is far from clear. One hypothesis could be that these two phenomena are unrelated, with the *pncr003;2L* performing two distinct functions, one in Ca^{2+} regulation and another in the maintenance of heart rhythm by a different mechanism, however such a hypothesis would still not explain that the same arrhythmic phenotype is produced by both the excess and lack of function of *pncr003;2L*. On the other hand, because there is evidence that the regulation of intracellular Ca^{2+} levels in vertebrate cardiomyocytes plays an important, and even essential, role in maintaining heart rhythm, by triggering

and regulating different ionic currents [86,87,88]. The hypothesis that in *Drosophila*, the concentration of Ca^{2+} needs to be tightly regulated in order to maintain the rhythmic contractions observed in wild-type hearts, seems more plausible. Indeed, there is also evidence in *Drosophila*, that disruption of Ca^{2+} levels in the heart produce arrhythmias [89]. In a way, the electrophysiological recordings of intracellular action potentials, provided by Jeremy Niven, which clearly show that individual mutant cardiomyocytes have action potentials with significant abnormalities compared to wild-type—or mutant flies expressing the rescue constructs—is in line with this hypothesis since the action potential of the cell is intrinsically linked to the ionic currents, which in vertebrate cardiomyocytes seem to be largely influenced by Ca^{2+} . Otherwise, these results provide another source of evidence, showing that the genotypes that present arrhythmias, also present a severe intracellular derangement which is likely to be linked with the observed misregulation of calcium.

Overall, it can be concluded that this work has characterised the *pncr003;2L* gene and the peptides it encodes, as regulators of cardiac Ca^{2+} cycling in *Drosophila melanogaster*, and has linked this misregulation with heart arrhythmias in the Fly. These results not only show that the *pncr003;2L* smORFs have an important function in flies, highlighting the importance of smORF genes in general, but also contribute by bringing forward the *Drosophila* adult heart, and this particular *pncr003;2L*-mediated Ca^{2+} regulation system, as a possible model to study the relation between Ca^{2+} misregulation and heart arrhythmias, and therefore to contribute to our understanding of heart disease.

Chapter VI - Identification of the *pncr003;2L* smORF as a functional homologue of the vertebrate Sarcolipin / Phospholamban family of regulators of the sarcoendoplasmic reticulum Ca^{2+} ATPase (SERCA).

1- Introduction:

I have shown, in the previous chapter of this thesis, that either the lack of function, or the over-expression, of the *pncr003;2L* peptides result in abnormal Ca^{2+} dynamics during heart contraction, and in heart arrhythmias. In accordance with the previous observation that the *pncr003;2L* peptides localise to the dyads, where the extrusion and uptake of Ca^{2+} during muscle contraction and relaxation occur, these phenotypes clearly indicate that *pncr003;2L* has an effect on the regulation of Ca^{2+} cycling in cardiac muscles. However, the molecular mechanism by which this regulation takes place remains to be determined.

The identification of the functional context of a gene could be achieved by means of a genetic interaction screen, in order to identify the genes that, in a haploinsufficient or over-expressed condition, interact with *pncr003;2L*, by suppressing or enhancing the *pncr003;2L* associated phenotypes. Before considering such a genetic screen for the characterisation of the functional context of *pncr003;2L*, and keeping in mind that this work should ultimately serve as a case study for the functional characterisation of smORFs, I explored the possibility of using an homology search approach, which could be applied in a more general way to other smORFs. This approach is based on standard, and modified BLAST searches, as well as on the use as a novel and powerful remote

homology search engine (PHYRE2) [30], which performs comparisons of predicted secondary structures, as well as sequence similarity, in order to search for other peptides, or protein domains, sharing some homology with the *pncr003;2L* peptides, and whose function may already be characterised. For such an approach, the very specific localisation and function of the *pncr003;2L* peptides identified so far, represent a way to narrow down the list of possible homologues to genes with similar characteristics. The identification of such homologues would provide an insight into the functional context of *pncr003;2L*, and the basis for experimental work, in order to validate the homology.

An extended homology search of this kind constitutes, by itself, an important part of the characterisation of the *pncr003;2L* gene, by determining the extent of conservation of this smORF gene, which so far, is only known to be conserved in *Drosophila pseudoobscura* [24].

In this chapter, through this extended sequence homology search, I describe the identification of the 30aa peptide encoded by *sarcolipin (sln)* as a possible homologue of *pncr003;2L*. Interestingly, *sln* acts as a regulator of the sarcoendoplasmic Ca^{2+} ATPase in vertebrate muscles. I provide evidence supporting this homology by identifying intermediate homologues between the human and *Drosophila* sequences, and use the functional assessment methods described in [Chapter V](#) to provide further evidence supporting the functional homology between the *pncr003;2L* peptides and the Sarcolipin / Phospholamban family of calcium regulators.

2- Results:

2.1-A BLAST search identifies homologue sequences for *pncr003;3L* in dipterans.

In the general introduction of this thesis, I have discussed how most bioinformatics methods, are generally ill-suited to dealing with the sequences of small peptides. These difficulties are exemplified by the use of basic local alignment tools (BLAST) in order to identify possible homologues for *pncr003;2L*. In the original paper describing *pncr003;2L* as non-coding [24], it was shown that a *pncr003;2L* homologue exists in the *Drosophila pseudoobscura* genome, because a probe specific to the *melanogaster* gene sequence hybridises with specific targets in southern blots and *in situ* hybridisation experiments in *pseudoobscura* tissues. However, no homologues for this smORF gene

were identified for its peptide sequences using a BLAST search with standard parameters, in which the peptide sequences are used in a search for translated nucleotides (tBLASTn), from EST databases that include the *Drosophila pseudoobscura* mRNA sequences (Figure 6.1A). In order to improve the detection of homologues with tBLASTn, I modified the search parameters to better suit the alignment of small sequences: I chose the PAM-30 substitution matrix instead of the standard Blosum-32 matrix, following the recommendations for sequences <35 aa long, provided by the provisional table of recommended substitution matrices from the NCBI BLAST help web-site (http://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html), and I relaxed the search parameters. This relaxation of the parameters was achieved by increasing the expected threshold of matches obtained purely by chance from 10 to 1000, and by removing compositional adjustment and low complexity region filters. With such optimised parameters it was possible to identify similar amino acid sequences for these peptides in most Drosophilids (including *Drosophila pseudoobscura*) and other dipterans, including other fly species (*Glossina moristans* and *Sarcophaga crassipalpis*) and mosquitoes (*Aedes Aegypti*, *Anopheles gambiae*, and *Armigeres subalbatus*) (Figure 6.1A). In all of these cases, the confidence of the sequence similarity between the *pncr003;2L* peptides and these hits is high (most having e-values of around $9e^{-06}$), and therefore homology between these sequences is quite likely. Furthermore, because the search was done in EST databases, it was possible to verify that the mRNA sequences of these hits did not contain other larger ORFs, and that the hits themselves corresponded to small ORFs of similar sizes to the *pncr003;2L* peptides, indicating that these are *bona fide* smORF homologues. Although these results expand the conservation of *pncr003;2L*, from what has so far been described (with the most distant homologue found before this search, being that in *Drosophila pseudoobscura*), none of the putatively homologous sequences identified have any functional annotations, which could help with the characterisation of the molecular function of the *pncr003;2L* peptides.

Figure 6.1: Initial analysis of sequence conservation and structure of the *pncr003;2L* peptides.

(A) Results of the tBLASTn searches, using the *pncr003;2L* peptides on EST databases, using either standard parameters (Blosum-32 matrix, Expected threshold of matches obtained purely by chance of 10, using compositional adjustment and low complexity region filters), or maximally relaxed parameters (PAM-30 matrix, Expected threshold of matches obtained purely by chance of 1000, removing compositional adjustment and low complexity region filters). The use of standard parameters identifies no homologues, whereas relaxed parameters identify homologues confined within the Dipterans. Amino acid colours reflect hydrophobicity (with red the most, and blue the least hydrophobic residues). (B) Different secondary structure prediction algorithms, from a web-base secondary structure prediction tool (<http://npsa-pbil.ibcp.fr>) predict an alpha-helical secondary structure for the *pncr003;2L* peptides. h: helical structure, c: random structure.

A

| Method | Result |
|---|---|
| BLAST standard parameters: | No homologues found |
| BLAST optimized parameters for small sequences: | <div>Identification of Dipteran homologues</div> <div><div>ORF 1</div><div><div>Melanogaster/1-28</div><div>MSEARNLFTTFFGILAILLFFFLYLIYAVL</div></div><div><div>Annanasse/1-28</div><div>MSEAKNLMTTFGILAVLLFFFLYIIYAVL</div></div><div><div>Mojavensis/1-28</div><div>MSEATNLTFTTFFGILAILLFFFLYIIYAVL</div></div><div><div>Wiiskonsin/1-28</div><div>MSEAKNLMTTFGILAILLFFFLYIIYAVL</div></div><div><div>Pseudoobscura/1-28</div><div>MSEAKNLMTTFGILAFLLFCLYLIYAVL</div></div><div><div>Sarcophaga/1-28</div><div>MSEAKSLSTFLLILAVLLSLLYLIYALF</div></div><div><div>Anopheles/1-33</div><div>MBETRNLMTTFFILIFLLLLLYLVYEFYOPEN</div></div><div><div>Armigeres/1-29</div><div>MBETKNLMTTFILIFLLLLLYVVKELIF</div></div><div><div>Aedes/1-27</div><div>MBETRNLMTTFLLIFLLFLLYLV--IF</div></div><div>***.***.***.***.***.***.***.</div></div> <div><div>ORF 2</div><div><div>Melanogaster/1-29</div><div>MNEAKSLFTTFLILAFLLFLLYAFYEAAF</div></div><div><div>Annanasse/1-29</div><div>MNEARSLFTTFLILAFLLFLLYAFYEAAF</div></div><div><div>Mojavensis/1-29</div><div>MNEARSLFTTFLILAFLLFLLYAFYEAAF</div></div><div><div>Wiiskonsin/1-29</div><div>MNEARSLFTTFLILAFLLFLLYAFYEAAF</div></div><div><div>Pseudoobscura/1-29</div><div>MNEARSLFTTFLILAFLLFLLYAFYEAAF</div></div><div><div>Glossina/1-25</div><div>MNEARSSFTTFLILAFLLFLLYTLV----</div></div><div><div>Sarcophaga/1-29</div><div>MNEAKSLFTTFVILVFLLSLLYIFYIAT</div></div><div>*****.*****.*****.*****.*****.</div></div> |

B

| | |
|-----------|-----------------------------------|
| ORF-A | MSEARNLFTTFFGILAILLFFFLYLIYAVL |
| DSC | hhhhhhhhhhhhhhhhhhhhhhhhhhhhhh |
| HNNC | chhhhhhhhhhhhhhhhhhhhhhhhhhhhhc |
| MLRC | ccchhhhhhhhhhhhhhhhhhhhhhhhhhhc |
| PHD | chhhhhhhhhhhhhhhhhhhhhhhhhhhhhc |
| SOPM | hhhhhhhhhhhhhhhhhhhhhhhhhhhhhh |
| Sec.Cons. | chhhhhhhhhhhhhhhhhhhhhhhhhhhhhc |
| ORF-B | MNEAKSLFTTFLILAFLLFLLYAFYEAAF |
| DSC | hhhhhhhhhhhhhhhhhhhhhhhhhhhhhh |
| HNNC | chhhhhhhhhhhhhhhhhhhhhhhhhhhhhc |
| MLRC | ccchhhhhhhhhhhhhhhhhhhhhhhhhhhccc |
| PHD | chhhhhhhhhhhhhhhhhhhhhhhhhhhhhc |
| SOPM | hctthhhhhhhhhhhhhhhhhhhhhhhhhhh |
| Sec.Cons. | chhhhhhhhhhhhhhhhhhhhhhhhhhhhhc |

Figure 6.1

2.2- The incorporation of secondary structure comparison, using the PHYRE2 homology search engine identifies the vertebrate *sarcolipin* smORF as a putative homologue for *pncr003;2L* .

Interestingly, the *pncr003;2L* peptides and their homologues have a highly hydrophobic amino acid constitution, rich in Phenylalanine (F), Isoleucine (I), Leucine (L), Alanine (A) and Valine (V) residues (Figure 6.1A). I tested whether such an evident bias in the nature of the amino acids constituting these peptides may convey a particular secondary structure, using a web-based bioinformatics secondary structure prediction tool (<http://npsa-pbil.ibcp.fr>). For each of the *pncr003;2L* peptides, this software produces an α -helical secondary structure prediction with all of the algorithms used (Figure 6.2B). Such a hydrophobic, α -helical structure, along with the 28 and 29 amino acid sizes of these peptides, is reminiscent of the transmembrane domains of membrane-bound proteins, which are often hydrophobic α -helical structures themselves, of about 20 amino acids. Such a structure would be consistent with the membrane bound subcellular localisation described so far for these peptides (Chapter III).




This prediction of a α -helical secondary structure for the *pncr003;2L* peptides allowed for a novel and powerful remote homology tool to be used in order to attempt the identification of more distant *pncr003;2L* homologues, which may have a characterised molecular function. This tool, called PHYRE2 (standing for protein homology/analogy recognition engine), uses a so called profile-profile algorithm, which incorporates secondary structure comparison to a conventional sequence similarity search, in order to identify possible remote homologous proteins, sharing similar secondary structure predictions and aa sequences [30]. This kind of algorithm, has been shown to outperform other powerful remote homology search methods such as the position-specific iterated (PSI)-BLAST method [90], which uses iterative homology searches that take into account, for each step, statistical calculations of mutational propensities at each position [91].

The PHYRE2 engine was used to identify possible homologues for one of the *pncr003;2L* peptides (*pncr003;2L* ORF A), and it yielded three hits (Figure 6.2A). One of these hits, corresponding to the Integrin α -m protein, only shares sequence similarity over a very small stretch of only 3 amino acids, and was therefore discarded. Another hit corresponds to one of the seven transmembrane domains of the Bacteriorhodopsin photoreceptor. Although such a match is interesting, because it is in line with the

transmembrane prediction discussed above, it is not very informative with regards to a possible functional homology at the molecular level, because no function for the Bacteriorhodopsin photoreceptor has been described in muscles, nor in calcium regulation. The third hit, however, is particularly interesting: First, unlike the other two hits, which correspond to large proteins of 1,200 and 300 aa, for the Integrin α -m and Bacteriorhodopsin proteins, respectively, this third putative homologue corresponds to another small, membrane peptide of 30 amino acids encoded by the human *sarcolipin* (*sln*) gene, which also happens to be a smORF gene (Figure 6.2A). Second, the *Sln* and *pncr003;2L* ORFA peptides have 7 identical residues, and 7 residues of similar nature over the 20 aa stretch highlighted as putatively homologous by PHYRE2, giving them an overall similarity score of 35%, which is higher than the 20% score obtained for the Bacteriorhodopsin protein. The sequence and structural similarities between the human *Sln* and Fruit fly *pncr003;2L* ORFA peptides, can be highlighted by a comparison of the structural diagrams that display the configuration for each of these peptides in the secondary structure either used, in the case of *Sln*, or generated, in the case of *pncr003;2L* ORFA, by PHYRE2 (Figure 6.2B). Indeed, such a diagram shows how, for this 20 aa stretch, the two peptides seem to be able to adopt remarkably similar structures, with identical, or very similar aa, in identical positions of the helix. Third, there is a striking similarity between the molecular function of *Sln* and the phenomenology so far described for the *pncr003;2L* peptides. Like *pncr003;2L*, *sln* is also a muscle specific gene, expressed exclusively in vertebrate somatic and cardiac muscles, where it has also been shown to regulate calcium cycling, via a direct physical interaction with the Sarco-endoplasmic reticulum Ca^{2+} ATPase (SERCA) [92,93,94,95]. Altogether, these pieces of evidence indicate that *pncr003;2L* could be the putative *Drosophila* homologue of the vertebrate *sln* gene family.

Figure 6.2 The PHYRE2, structural and sequence homology search engine identifies Sln as a *pncr002;2L* homologue. (A) Results of a PHYRE2 search on EST databases, querying the *pncr003;2L* ORFA peptide. This search yielded three hits (Figure 6.2 A): One of these hits corresponds to the Integrin α -m protein, but only shares sequence similarity over a stretch of only 3 amino acids, another hit corresponds to one of the seven transmembrane domains of the Bacteriorhodopsin photoreceptor. And a third hit, with the best similarity score, corresponds to the human Sln 30 aa long peptide involved in calcium regulation in muscles. The colours in the aa alignments represent the CLUSTALW standard colours, regions with alpha-helical predictions are indicated by green helices. (B) Structural display of the aa configuration of the 20 aa stretches identified by PHYRE2 as homologous, showing that the *pncr002;2L* peptide and Sln could adopt very similar secondary structures.

A

| Method | Result | | | |
|--|-------------------------------------|---|------------|--|
| PHYRE2 structural homology search engine: | Identification of a human homologue | | | |
| | Template | 3D Model | Confidence | % Id. Template Information |
| | c1bcta_ |  | 29.9 | 20 Bacteriorhodopsin, photoreceptor |
| | c1bda_ |  | 9.7 | 35 Sarcophilin, membrane protein |
| | c2kda_ |  | 8.0 | 80 Integrin alpha-m, cell adhesion protein |
| <div><div>Predicted Secondary structure</div><div>Query Sequence</div><div>Template Sequence</div><div>Template Known Secondary structure</div></div> <div><div>2 10 20</div><div>MMSEARNLFTTFGILAILLFFLYLI</div><div>MRPEVASTFKVLRNVTVLWSAYPV</div><div>163 170 180</div></div> | | | | |
| <div><div>Predicted Secondary structure</div><div>Query Sequence</div><div>Template Sequence</div><div>Template Known Secondary structure</div></div> <div><div>7 10 20</div><div>RNLFITTEGLAILLFFLYLI</div><div>RELFLNFTIVLITVILMWLL</div><div>5 10 20</div></div> | | | | |
| <div><div>Predicted Secondary structure</div><div>Query Sequence</div><div>Template Sequence</div><div>Template Known Secondary structure</div></div> <div><div>2</div><div>MMSEIA</div><div>MMSEIG</div><div>12</div></div> | | | | |

B

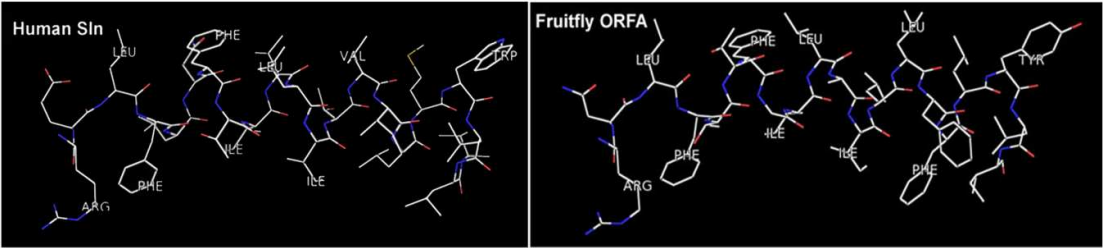


Figure 6.2

2.3- A BLAST search using a phylogenetic consensus sequence between Dipteran *pncr003;2L* sequences and human Sarcopin identifies homologues throughout the arthropoda phylum.

In light of the striking similarities between these peptides, it could be argued that *pncr003;2L* and *sln* may belong to the same family of smORFs, having a function in calcium regulation which has been conserved across evolution, and which has so far been hidden by 1) the inaccurate annotation of *pncr003;2L* as non-coding, and 2) the inability of standard homology search tools to identify distant homologues with small sequences. If this homology is real, one would expect to find homologous peptides in other intermediate species. However a BLAST search using the *pncr003;2L* sequences as a query has already been shown, despite the most optimised parameters, to produce hits that remain confined within the dipteran order.

I reasoned that if the homology between *pncr003;2L* and *sln* is real, then a consensus sequence, obtained from an alignment of the different Dipteran *pncr003;2L* peptides and the human *Sln* peptide (Figure 6.3A), favouring the residues conserved between any of the members of the Diptera order, and the human species may produce hits corresponding to the sequences of intermediate species. This is indeed the case, as a tBLASTn search using such a consensus sequence, and the same optimised parameters as those used for the original tBLASTn search, yield hits across the arthropoda phylum, including sequences belonging to members of the hexapoda, crustacea, arachnida and xiphosura sub-phylums (Figure 6.3A). As with the original BLAST search, I verified that the mRNA sequences encoding each of these hits do not code for any longer ORFs, thereby confirming that these hits also correspond to smORF genes. Interestingly, the polycistronic arrangement observed in *Drosophilids*, and other Dipterans, such as *Sarcophaga crassipalpis*, can also be observed in other examples of these newly identified arthropods, such as *Bombyx mori* and *Ixodes scapularis*, whose transcripts contain multiple *pncr003;2L*-like ORFs (Figure 6.3A).

The relationship between the *pncr003;2L* and *Sln* peptides is further supported by the fact that sequences of the more basal arthropods, such as those of arachnids (*Ixodes scapularis*) are similar enough to the vertebrate sequences to produce hits, upon the same tBLASTn search as that performed above, corresponding to *Sln* in basal vertebrate species, and also corresponding to the 52 aa peptide encoded by *phospholamban (pln)*, which is a paralogue of *sln*, also known to regulate calcium

cycling during muscle contraction in skeletal and cardiac muscles through a direct physical interaction with SERCA [96].

Unlike Sln, which is composed of a small N-terminal luminal domain 7 aa long, a transmembrane domain, and a small C-terminal cytosolic domain 5 aa long, Pln has a relatively large N-terminal cytoplasmic domain of 30aa, and a transmembrane domain, which is related to that of Sln. An alignment of the *pncr003;2L* Dipteran peptides, Sln, and the transmembrane domain of the human Pln, reveals the patterns of conservation in this family of peptides. In fact, several amino acid changes appear to be semi-conservative between the Pln / Sln and *pncr003;2L* sequences, like in the cases of the phenylalanine (F)/ (Tyrosine)Y, and (Tryptophan)W/(Tyrosine)Y, which represent a conservation in aromatic residues, in positions 37 and 48, respectively, and likewise, most of the hydrophobic aa often alternate between Isoleucine (I) and Leucine (L) in different arthropod and vertebrate species. There also seems to be a prevalence of Serine (S) and Threonine (T) residues in the N-terminus of all of these peptides, as well as negatively charged Glutamic acid residues (E). Interestingly, the sequence of the basal arthropod *Ixodes_A* peptide shows, a greater extent of sequence conservation when compared to a basal vertebrate sequence (*Danio rerio*), with 13 out of the 32 aa in the *Ixodes_A* peptide being identical, and 4 others of similar nature to those of *Danio_Pln*, and a conservation score of 21% as calculated by ClustalW, (compared to the score of 17% for the human Pln and *Drosophila pncr003;2L* alignment) (Figure 6.3B). These results, which show patterns of conservation between the arthropod and vertebrate sequences, as well as higher conservation between basal arthropods and basal vertebrates suggest that the *sln/pln* and *pncr003;2L* genes are part of the same family of smORFs, which appears to be conserved across evolution, with its origin predating the last common ancestor of these phyla, the Urbilaterian (Figure 6.4).

Figure 6.3: A tBLASTn search using the phylogenetic consensus between the pncr003;2L peptides and Sln, identifies intermediate homologues. (A) The phylogenetic consensus sequence (ETRSLFTTFXILAILLFLLWLLYE) between the Dipteran pncr003;2L peptides and Sln, obtained by favouring the residues conserved in the Dipteran and human sequences (underlined residues), yield intermediate homologous sequences throughout arthropods when used as a query on a tBLASTn search with maximally relaxed parameters (PAM-30 matrix, Expected threshold of matches obtained purely by chance of 1000, removing compositional adjustment and low complexity region filters). The grey line delimits the hits identified by this consensus tBLASTn search, homologues from the craniata subphylum can be identified by a tBLASTn search querying for the basal arthropod sequences (crustacea / arachnida). (B) The sequence of pncr003;2L from basal arthropods, such as Ixodes A, and Ixodes B share greater similarity to basal vertebrate sequences, such as those of *Danio rerio* (zebra fish) as shown by the alignment between Ixodes A and Danio_Pln, and Ixodes B and Danio_Sln, which shows more conservation than the alignments of the *Drosophila* peptides to either human_Pln or human_Sln, and producing vertebrate hits when using them as queries with the same tBLASTn search as in (A). Identical residues (*) and conservative residues (:) are indicated.

A



B

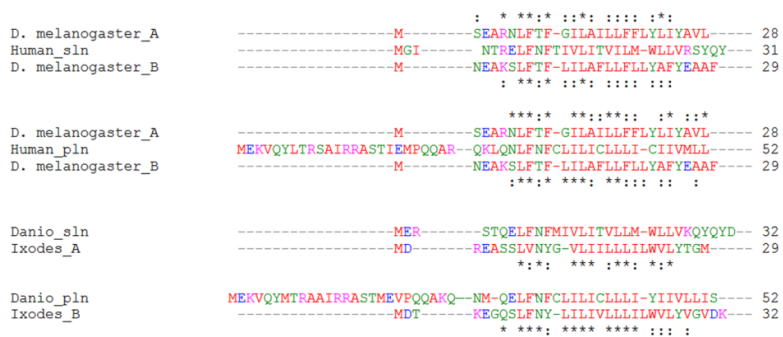


Figure 6.3

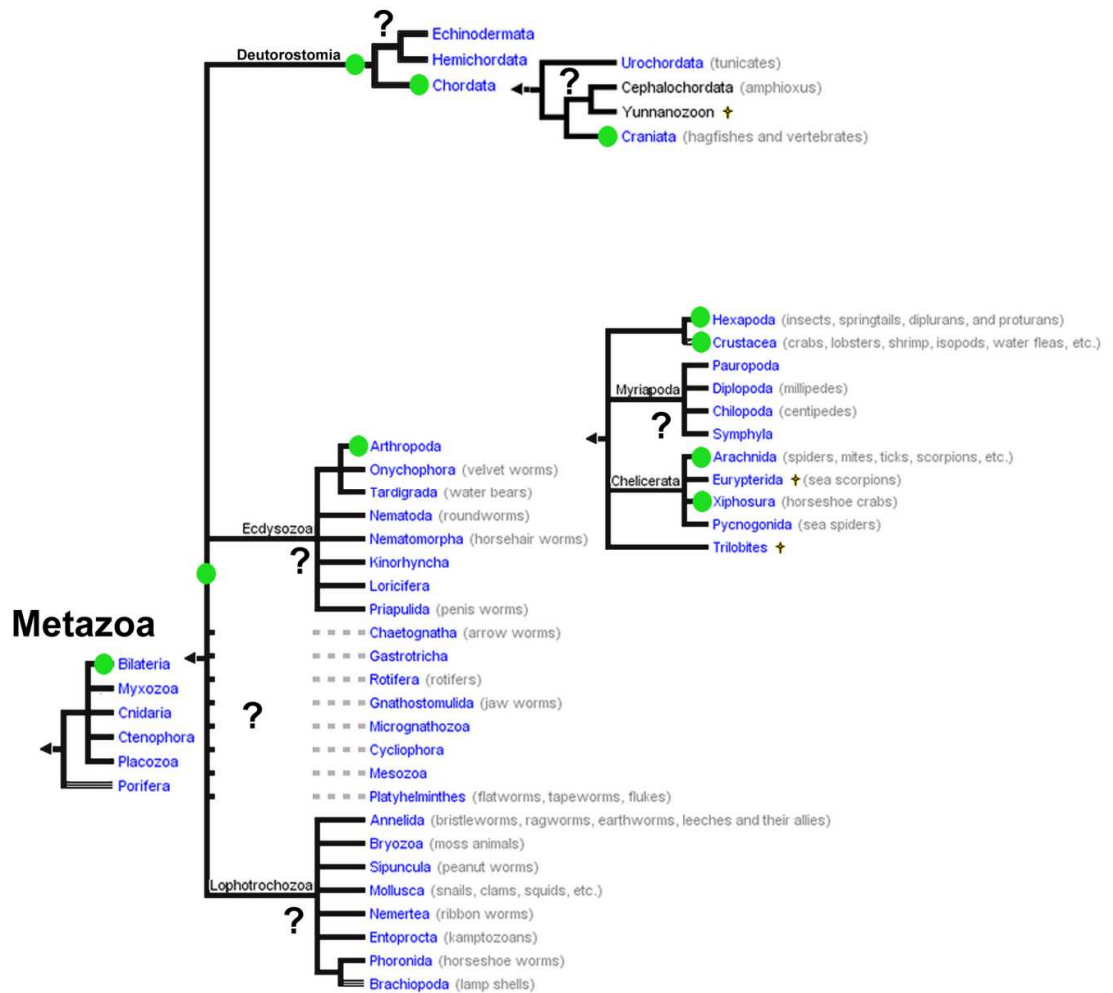


Figure 6.4

Figure 6.4: The Sln and pncr003;2L smORFs seem to have a common bilaterian ancestor.

(A) Phylogenetic tree showing the different taxonomic ranks where the PIn /PIn and pncr003;2L peptides have been identified by the homology searches performed in this work (Green circles), pointing to a possible common Bilaterian common ancestor. Question marks highlight the taxonomic ranks where no evidence for homologues of these

2.4- The *pncr003;2L* peptides co-localise, and interact genetically with *Ca-P60A*, the *Drosophila melanogaster* homologue of *SERCA*.

The identification of the *sln /pln* smORF genes as possible homologues for *pncr003;2L*, would provide a specific molecular context for the function of the *pncr003;2L* peptides. Both *Sln* and *Pln* have been shown to negatively regulate the activity of *SERCA*, by binding to a site in the transmembrane domain of the Ca^{2+} ATPase. *SERCA* is a highly conserved gene, present in both animal and plant cells [97]. Although in vertebrates three *SERCA* paralogues have been identified, known as *SERCA-1,2* and *3* —all of which can be regulated by the *Sln /Pln* peptides—, in *Drosophila melanogaster* a single gene, annotated as *Ca-P60A*, has been identified as a *SERCA* homologue [98]. Importantly, the *Drosophila Ca-P60A* gene, which shares 72% of identity with its vertebrate homologue [98], has been shown to regulate calcium cycling in *Drosophila* neurones, and in somatic and cardiac muscles [35,89].

In order to assess the possible interaction between the *pncr003;2L* peptides and the *Ca-P60A* pump, I used a primary antibody specific to *Ca-P60A*, kindly provided by S. Sanyal [35], to study the localisation of the *Ca-P60A* pump in relation to the *pncr003;2L* peptides in IFMs and cardiac muscles. These *Ca-P60A* antibody stainings, which were performed in a genetic background expressing the *pncr003;2L ORFA-GFP*, or N-terminal FLAG-Hemagglutinin tagged construct *pncr003;2L FH-ORFA*, show that the endogenous *Ca-P60A* pump co-localises perfectly with the *pncr003;2L* peptides, in the dyads and perinuclear membrane of IFMs (Figure 6.5A) and cardiomyocytes (Figure 6.5B). Although this co-localisation between the *pncr003;2L* peptides and *Ca-P60A* is a pre-requisite for a protein-protein interaction to occur between these two entities, and is therefore in agreement with such interaction, it does not prove that this interaction actually occurs. The physical interaction between the *pncr003;2L* peptides and *Ca-P60A* is supported by the work of J.I. Pueyo, and F.M.G. Pearl, performed in parallel to the work presented here (12), who performed biochemical and bioinformatics assays, respectively, both supporting the physical interaction between these entities. Their work is addressed in more detail in the discussion of this chapter.

The evidence presented so far in this work —and in the work of J.I. Pueyo and F.M.G. Pearl—, strongly supports the homology between the *pncr003;2L*, and the *sarcolipin* and *phospholamban* genes. In order to reflect this homology, and because *pncr003;2L* is

in fact a coding gene and not a non-coding RNA, this smORF gene was renamed *sarcolamban (scl)*.

I have so far shown that the lack of function of *scl*, in a *Df pncr003;2L* (or *Df scl*) background produces heart contractions which are arrhythmic, and during which the calcium transients have larger amplitudes than wild type. If the function of *scl* is indeed homologous to that of *pln* and *sln*, which have been thoroughly proven to act as inhibitors of SERCA in vertebrates [92,94,95,96,99,100,101], it could then be stated that the arrhythmicity and abnormal calcium transients observed in *scl* null flies would be due to the release of the inhibition of Ca-P60A by the Scl peptides.

In order to assess whether *Ca-P60A* has an effect on these arrhythmicity and calcium transient phenotypes, I performed a genetic interaction assay between *slc* and *Ca-P60A*, taking advantage of a *Ca-P60A* homozygous lethal allele (*Ca-P60A*^{Kum 295}) generated by Sanyal et.al. [35]. For this assay, the *Ca-P60A*^{Kum 295} allele was introduced, as heterozygous, into the *Df scl* background. This genetic condition (*Df γ-ray 6*, *Ca-P60A*^{Kum 295} / *Df(2L) 12, +*) was achieved by placing over the deficiency *Df(2L) 12*, a recombinant chromosome in which the *Ca-P60A*^{Kum 295} allele was linked with the *Df γ-ray 6*. This recombinant chromosome was obtained by screening the F1 progeny of *b Df γ-ray 6 sp* / *Ca-P60A*^{Kum 295} females for the loss of the *sp* marker and retention of the *b* marker, and by testing the resulting recombinants for lethality over both *Df γ-ray 6* and *Ca-P60A*^{Kum 295}. Interestingly, the heart arrhythmicity is significantly corrected in *Df γ-ray 6*, *Ca-P60A*^{Kum 295} / *Df(2L) 12, +* flies compared to *Df scl* flies (*Df γ-ray 6*, + / *Df(2L) 12, +*) (Figure 6.6A and B). Similarly, the calcium transients of *Df scl* hearts have significantly larger amplitudes than those of *Df γ-ray 6*, *Ca-P60A*^{Kum 295} / *Df(2L) 12, +* hearts, which have the same amplitudes as wild-type controls (Figure 6.6C). These results indicate that the hemizygous condition of *Ca-P60A* is able to rescue the arrhythmicity phenotype, as well as the abnormal calcium transient amplitudes observed in *Df scl* hearts. This correction of the *scl* null phenotypes, by the reduction of the *Ca-P60A* genetic dosage, is in line with an inhibitory role of the Scl peptides on the Ca-P60A enzyme, and most importantly, proves that *scl* and *Ca-P60A* are functionally linked.

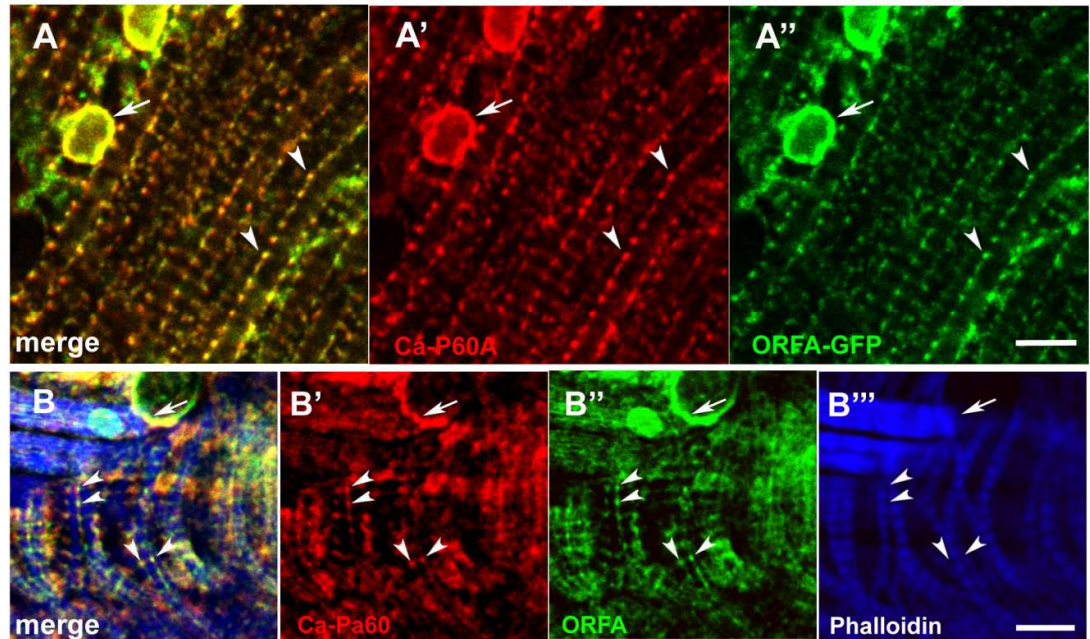


Figure 6.5

Figure 6.5: The pncr003;2L peptides co-localise with Ca-P60A, the *Drosophila* SERCA homologue, in the dyads. (A-A'') Confocal microscopy images showing the co-localisation of the pncr003;2L-GFP tagged peptides (green) and Ca-P60A SERCA (red) in the SER and dyads (arrowheads) surrounding the sarcomeres of indirect flight muscle myofibrils. (B-B'') Confocal microscopy images showing the co-localisation of the pncr003;2L FH-ORFA tagged peptides (green) and Ca-P60A SERCA (red) in the SER and dyads (arrowheads) surrounding the sarcomeres (blue, phalloidin) of adult cardiomyocytes. Scale bars: (A-A'') = 10 μ m; (B-B'') = 5 μ m.

Figure 6.6: *pncr003;2L* interacts genetically with *Ca-P60A*. (A) Kymographs representing the pattern of heart contractions in *Df(2L)scl* (*Df γ-ray 6, + / Df(2L) 12, +*) flies, showing an arrhythmic pattern of heart contractions, and *Df(2L)scl* carrying a *Ca-P60A* null allele (*Df γ-ray 6, Ca-P60A^{Kum 295} / Df(2L) 12, +*), showing a regular pattern of heart contractions. (B) A quantification of the arrhythmicity index between these genotypes shows that the *Df γ-ray 6, Ca-P60A^{Kum 295} / Df(2L) 12, +* flies have a significantly lower arrhythmicity than *pncr003;2L* null hearts, as determined by a two-tailed Mann Whitney test, (U=22, p<0.003,). n=15-20 flies per genotype. (C) Averaged G-Camp3 fluorescence signals of calcium transients normalised as in Figure 5.7B, and plotted in relation to the wild-type average maximum signal. *Df γ-ray 6, + / Df(2L) 12, +* hearts show significantly higher calcium transient amplitudes than *Df γ-ray 6, Ca-P60A^{Kum 295} / Df(2L) 12, +* hearts, as determined by a two-tailed Mann Whitney statistical test, (U=5, p=0.0017), n=10 flies per genotype.

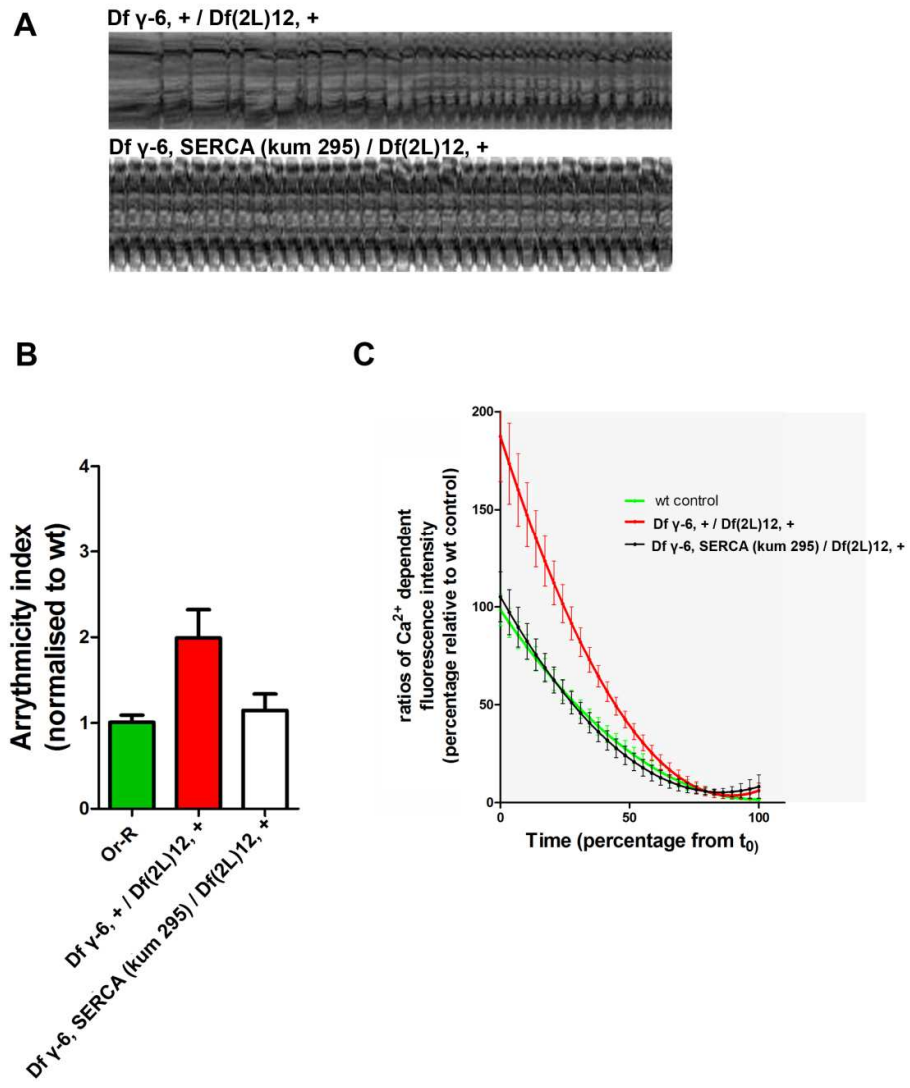


Figure 6.6

2.5- The vertebrate Sln and Pln peptides partially recapitulate the function of the Scl peptides in flies.

In order to further assess the homology between *sln*, *pln*, and *scl* at the functional level, I tested the functional equivalence between the vertebrate and *Drosophila* peptides within the context of *Drosophila melanogaster*. For this, different constructs were generated, and transfected into flies, in order to generate transgenic *Drosophila* lines expressing the vertebrate peptides.

First, in order to compare the subcellular localisation of the vertebrate and *Drosophila* peptides, transgenic lines carrying N-terminal FLAG-Hemagglutinin tagged Sln (FH-Sln) and Pln (FH-Pln) constructs were generated. These N-terminal tagged constructs were obtained by cloning the Sln or Pln ORFs into the same N-terminal FLAG-hemagglutinin tag vector as that used for the FH-ORFA and FH-ORFB constructs. The FH-Sln and FH-Pln constructs were co-expressed with *scl* ORFA-GFP in muscles — driven with the *Dmef2-GaL4* driver— and their subcellular localisation assessed by immunohistochemistry (Figure 6.7). These double staining experiments show a perfect co-localisation between the Sln and Scl ORFA peptides (Figure 6.7A), and between the Pln and Scl ORFA peptides (Figure 6.7B) in the dyads and perinuclear membrane of IFM, which, is in agreement with the hypothesis that all of these peptides interact with the same protein (SERCA/Ca-P60a).

Second, in order to compare the function of the Sln, Pln, and Scl peptides, I tested the ability of the vertebrate peptides to induce similar over-expression phenotypes as those obtained by over-expressing *scl*, or to rescue the *scl* null condition. For this, transgenic constructs were generated, in which the *scl* ORF was substituted by either the *sln* or *pln* ORF, in the different *scl* constructs used for the rescue or over-expression experiments presented in [Chapter V](#). This ORF substitution strategy was implemented in order to ensure that the expression context of the vertebrate peptides is as similar as possible to that of the Scl peptides.

The over-expression of these constructs in a wild-type background, like the over-expression of Scl itself, leads to an increase in the arrhythmicity index (Figure 6.8A). In the case of Sln, the increase is very similar to that observed with the Scl peptides, compared to the over-expression of the *pncr003;2L FS* control, or to the control

experiments in which each of the UAS lines were present, but not the *Dmef2-GaL4* driver. The over expression of Pln, on the other hand, produces a more pronounced arrhythmicity, which is almost two fold in comparison with that produced by either the Sln or the Scl peptides. Interestingly, the over-expression of either Sln or Pln leads to a similar reduction in the amplitude of the calcium transients as that observed with the Scl peptides (Figure 6.8B), indicating that in this over-expression conditions, the vertebrate Sln and Pln peptides affect the activity of SERCA in a similar way as the over-expression of Scl (Figure 5.8)

When either of the constructs carrying the Sln or Pln peptides, are expressed in the *Df scl* background, using the *Dmef2-GaL4* driver, a small correction in the arrhythmicity index can be observed compared with *Df Scl* flies expressing the *pncr003;2L FS* control construct, however neither of these changes are statistically significant (Figure 6.8C). Interestingly, the expression of Sln leads to a very minor reduction in the amplitude of the calcium transients of *Df scl* mutants, compared to the expression of the *pncr003;2L FS* control, which is not statistically significant (Figure 6.8D), and is therefore consistent with the minor effect that this particular vertebrate peptide has in the arrhythmicity of the *Df scl* mutant hearts. On the other hand, the expression of Pln in the *Df scl* mutant background, leads to a significant reduction in the calcium transient amplitude, compared with *Df scl* hearts expressing the *pncr003;2L FS* control (Figure 6.8D), which is comparable with the reduction in amplitude induced by the expression of *scl* in this same mutant background.

Regarding Sln, it seems as though the effect of this particular peptide could be synergistic with the inhibition of Ca-P60A by Scl, to produce the observed arrhythmicity and reduction in calcium transients upon its expression in a wild-type background, where Scl is normally expressed. However, the effect of Sln on its own seems insufficient to rescue the mutant phenotypes induced by the loss of function of Scl.

In the case of Pln, it is interesting that its effects on calcium are similar to those of Scl, in both the rescue and over-expression conditions. There is however a discrepancy between the inability of this particular vertebrate peptide to fully rescue the arrhythmic phenotype, and its capacity to restore the calcium transients, which is almost as efficient as that of *scl* itself.

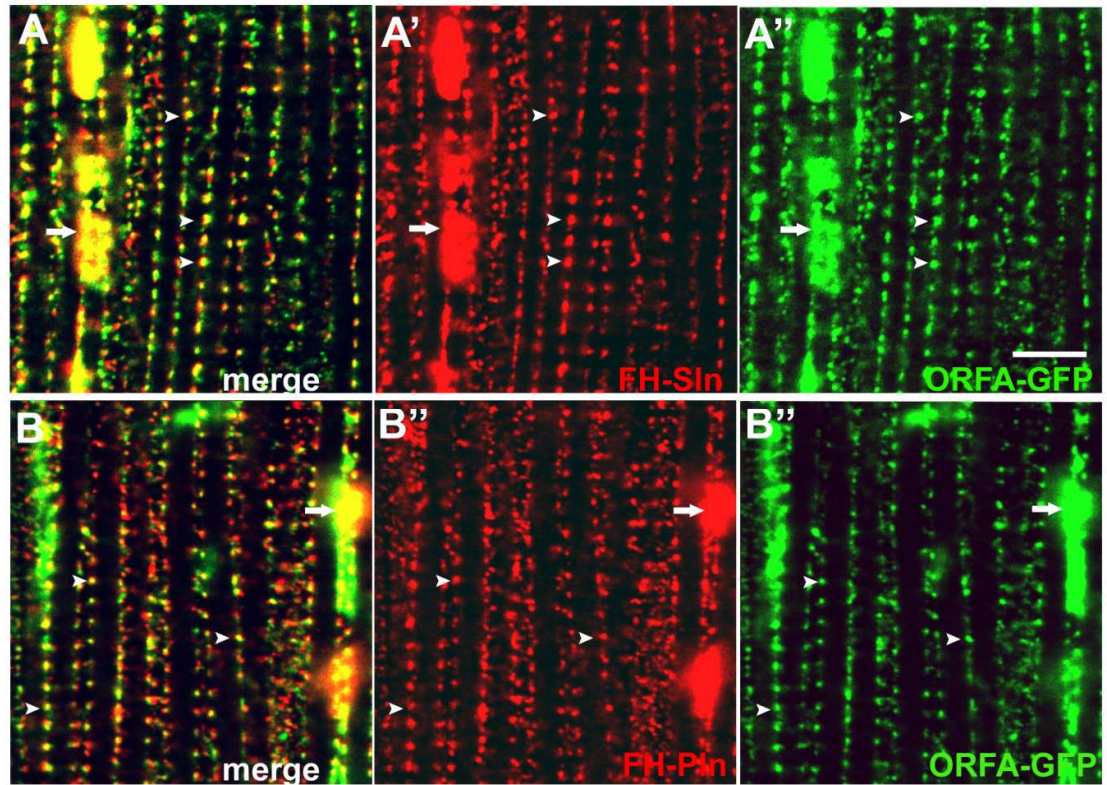


Figure 6.7

Figure 6.7: The pncr003;2L peptides co-localise with Sln and Pln in the dyads.

(A-B) Confocal microscopy images showing the co-localisation of the pncr003;2L-GFP tagged peptides (green) and (A) N-terminal FLAG- Hemagglutinin tagged Sln (FH-Sln) (red), or (B) N-terminal FLAG-Hemagglutinin Pln (FH-Pln) (red), in the perinuclear SER (arrows) and dyads (arrowheads) surrounding the sarcomeres of indirect flight muscle myofibrils. Scale bar: (A-B'')=10 μ m.

Figure 6.8: Vertebrate Sln and Pln peptides partially recapitulate the function of the Scl peptides.

(A) Quantification of the arrhythmicity induced by the over-expression of the vertebrate Sln and Pln peptides, compared to the *Drosophila pncr003;2L* ORFA and ORF B peptides. When either Sln or Pln are over-expressed in a wild-type background (*Dmef2>Sln*, and *Dmef2>Pln*) they induce a significant increase in arrhythmicity index, compared to the over expression of the *pncr003;2L FS* control construct, as determined by a two-tailed Mann Whitney statistical test, (U=125, p<0.0005) and (U=108, p<0.0005), respectively. Notice the increase in arrhythmicity by the over-expression of Sln is similar to that previously observed with the *Drosophila* peptides, while the increase in arrhythmicity induced by the over-expression of Pln, is greater compared to the other conditions. The expression of the Pln or Sln constructs, without the *Dmef-GaL4* driver has no effect on arrhythmicity. n=10 flies per genotype.

(B) Averaged GCaMP3 fluorescence signals of calcium transients normalised as in Figure 5.7B, and plotted in relation to the wild-type average maximum signal. Hearts over-expressing the Pln construct (*Dmef2>Pln*) show significantly reduced calcium transient amplitudes compared to hearts expressing the *pncr003;2L FS* control (*Dmef2>pncr003;2L FS*) as determined by a two-tailed Mann Whitney statistical test, (U=11, p=0.05). The hearts of flies expressing the Sln construct (*Dmef2>Sln*) show a smaller reduction in calcium transient amplitudes compared to *Df(2L)Scl* hearts expressing the *pncr003;2L FS* (U=12, p=0.019). n=8-10 flies per genotype.

(C) A quantification of the arrhythmicity of *pncr003;2L* mutant conditions (*Df(2L)scl*), in which the vertebrate Sln or Pln peptides were expressed to test their capacity to perform a phenotypical rescue. *Df(2L)scl* mutants expressing either Sln or Pln (*Df(2L)scl*, *Dmef2>Sln*, and *Df(2L)scl*, *Dmef2>Pln*) show a small reduction in arrhythmicity index compared to *Df(2L)scl* mutant hearts, or *Df(2L)scl* expressing the *pncr003;2L FS* control, these reductions however were not significantly different. n=15-20 flies per genotype.

(D Averaged G

CaMP3 fluorescence signals of calcium transients normalised as in Figure 5.7B, and plotted in relation to the wild-type average maximum signal. *Df(2L)scl* hearts

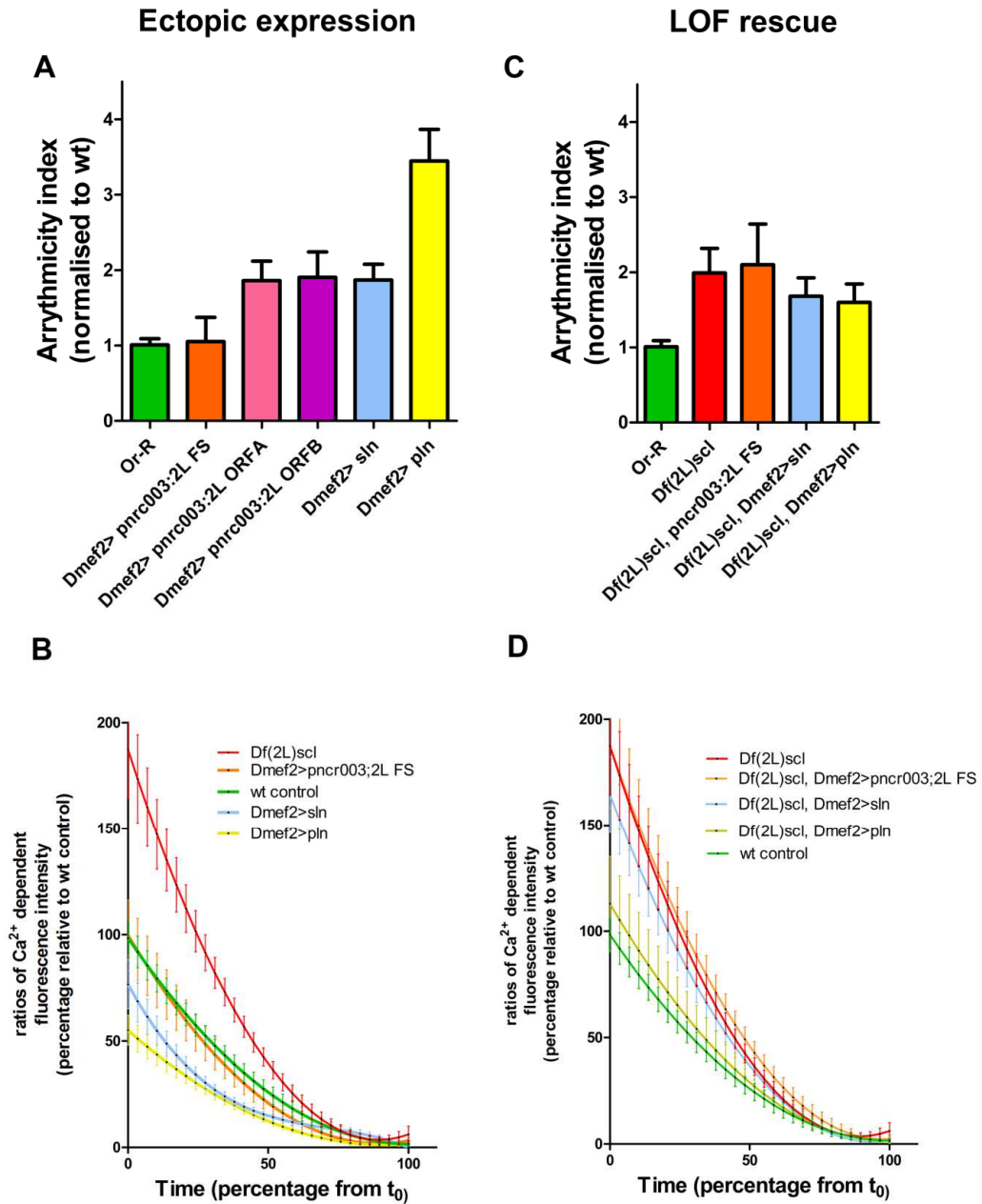


Figure 6.8

expressing the Pln construct (*Df(2L)scl*, *Dmef2>Pln*) show significantly reduced calcium transient amplitudes compared to *Df(2L)scl* hearts expressing the *pncr003;2L FS* control (*Df(2L)scl*, *Dmef2>pncr003;2L FS*) as determined by a two-tailed Mann Whitney statistical test, ($U=4, P<0.005$). *Df(2L)scl* hearts expressing the Sln construct (*Df(2L)scl*, *Dmef2>Sln*) show a very small reduction in calcium transient amplitudes compared to *Df(2L)scl* hearts expressing the *pncr003;2L FS* construct, which is not statistically significant. $n=8-10$ flies per genotype.

3- Discussion.

3.1 A phylogenetic analysis supports the homology between Scl and Pln/Sln

The work presented in this chapter proposes a molecular function for the regulation of *pncr003;2L* on Ca^{2+} cycling during cardiac muscle contraction by identifying the vertebrate *sln/pln* gene family, which are the main regulators of the Sarco-endoplasmic reticulum Ca^{2+} ATPase, as functional homologues of *pncr003;2L*. This homology was initially identified by the PHYRE2 homology search engine, and is supported by the existence of intermediate homologue sequences, which were identified in a tBLASTn search, using the phylogenetic consensus sequence between the Dipteran *pncr003;2L* peptides and human Sln.

Although homologous sequences were not identified in all intermediate phylogenetic ranks between the arthropoda and craniata sub-phyla, the homology between the vertebrate and arthropod sequences is still favoured by the higher conservation of the more basal arthropod and vertebrate sequences, and by the prevalence of semi-conservative amino acid changes between the Pln / Sln and *pncr003;2L* (Scl) sequences. Further supporting this homology, the work of J.P. Couso [33], who performed an extended phylogenetic analysis of these sequences, shows that it is possible to reconstitute, using the Sln, Pln and Scl sequences, and an unrelated control sequence of similar size, a phylogenetic tree which clusters the Sln, Pln and Scl groups together, while effectively out-grouping the unrelated sequence, and which accurately reconstitutes the phylogenetic distances between these sequences.

The lack of homologues in the other intermediate phylogenetic ranks, could be explained by the loss or divergence of these regulatory peptides, which may have either only been maintained in arthropods and vertebrates, or which may still be present in the other ranks, but with divergence to an extent that the BLAST search fails to identify them, similarly as when a query with the Scl sequence fails to identify Sln.

Alternatively, it is possible that other intermediate species where these sequences have been conserved do not have enough coverage in their EST libraries to allow the identification of these sequences. If it is the case that this family of genes is conserved in other intermediate species, the ongoing generation of higher quality sequence libraries for these intermediate phyla and the development of increasingly powerful methods for remote homology identification, will certainly lead to their identification in the near future.

3.2 Evidence supporting the physical interaction between Scl and Ca-P60A

Regarding the functional relation between *scl* and *Ca-P60A*, I firstly show that the Scl tagged peptides, whose subcellular localisation has been proven to be real throughout this work by different observations such as the lack of such patterns in other membrane bound markers, or the ability of these tagged peptides to rescue the *Scl* null phenotypes, co-localise extensively with the endogenous *Ca-P60A* in the dyads and perinuclear region of IFMs and cardiac myocytes. As stated in the results section of this chapter, this co-localisation is a pre-requisite for a protein-protein interaction to take place between Scl and Ca-p60A, but does not prove it. The physical interaction between the *pncr003;2L* peptides and *Ca-P60A* is supported by the work of J.I. Pueyo, and F.M.G. Pearl [33], who parallel to the work presented here, performed biochemical and bioinformatics assays, respectively, with both supporting the physical interaction between these entities. J.I. Pueyo performed an immunoprecipitation assay, in *Drosophila* S2R+ cells co-expressing either of the N-terminal Hemagglutinin-FLAG tagged *pncr003;2L* ORFA (FH-ORFA), *pncr003;2L* ORFB (FH-ORFB), Sln (FH-Sln) and Pln (FH-Pln) peptides, and Ca-P60A. In this assays, it was shown that both *Drosophila* peptides, and both vertebrate peptides, co-localise with Ca-P60A in S2R+ cells and were able to co-immunoprecipitate Ca-P60A, showing that the *pncr003;2L* peptides, as well as the vertebrate peptides, are able to bind Ca-p60A.

This binding is further supported by the work of F.M.G Pearl, who modelled computationally the docking of the *pncr003;2L* ORF A and ORFB peptides onto Ca-P60A, after building a structural model for the Scl and Ca-P60A protein structures by threading their sequences onto the secondary structures of Pln, Sln, and SERCA. Importantly, these vertebrate structures were all obtained from an X-ray crystallography structural model of the Sln-Serca complex [102], which also served to guide the virtual dockings. The calculated binding energies obtained for the Scl peptides were very similar to those obtained for the Sln and Pln peptides, with their respective Ca-P60A or SERCA targets (See Annex 4). The Scl peptides are in fact predicted to be able to dock into the same transmembrane pocket, as Sln and Pln, with all of these peptides producing similar predicted energy shifts in the overall structure, which essentially indicates that the *Drosophila* and vertebrate peptides have comparable properties with respect to their binding their respective Ca-P60A or SERCA targets. Furthermore, the results of F.M.G Pearl also show that the key residues for the Sln and Pln binding to

SERCA, determined by the crystal structures and by the bioinformatics analysis, are conserved in the *pncr003;2L* peptides.

3.3 Experimental evidence supporting the functional relation between Scl and Ca-P60A, and the functional homology between Scl Pln/Sln

In agreement with a physical interaction between Scl and Ca-P60A and the functional homology between Scl and the Sln/Pln peptides, the work presented here provides evidence of a functional interaction between Scl and Ca-P60A through the genetic interaction observed between these two genes. The results of this genetic interaction, in which the hemizygous condition of Ca-P60A rescues the abnormal calcium transients and the arrhythmicity of *scl* null flies, are in line with the peptides encoded by this gene having an inhibitory effect on Ca-P60A. These results fit with a model where the *scl* null condition would lift the Scl-mediated inhibition of Ca-P60A, leading to a constantly up-regulated enzymatic activity for this Ca^{2+} ATPase, which would be responsible for the abnormal Ca^{2+} transients and arrhythmia phenotypes. In such a model, the reduction of the genetic dosage for *Ca-P60* would indeed be expected to compensate for its excessive activity in the absence of Scl.

Regarding the effects of Scl on the Ca^{2+} transients, and its inhibitory relation with Ca-P60A, it needs to be highlighted that different studies, which have focused on the loss or excess of function of the Sln and Pln peptides in vertebrate cardiomyocytes, show changes in Ca^{2+} transient amplitudes in these conditions, which are remarkably similar to those presented in this work (See Annex 5) [99,103], with the lack of function of the vertebrate peptides producing calcium transients with higher amplitudes, and their excess of function, transients with lower amplitude, than wild type.

The model to explain these variations in calcium transients in vertebrates, is that the enhanced activity of the SERCA pump, in Pln or Sln null conditions, is more effective in replenishing the sarco-endoplasmic reticulum (SER) with Ca^{2+} , leading to higher concentrations of luminal calcium in this organelle. Because the rates of calcium release through the activated RyR, are regulated by the concentration of luminal Ca^{2+} and the intracellular cytoplasm [104,105], this higher luminal concentration leads to higher amounts of calcium being released, and to the higher amplitude in calcium transient. In over-expression conditions the opposite occurs, leading to lower amounts of Ca^{2+} stored and released by the SER, and to Ca^{2+} transients with lower amplitudes. Evidence supporting this model has recently been provided in a study which shows that both the

SERCA Ca^{2+} uptake and the SER calcium content, which was quantified by caffeine-induced SER depletion, are both enhanced in Sln and Pln double knock-down mice, with the mutant calcium transients being proportional to this elevated content [106]. Since the lack of function or over-expression of Scl in *Drosophila* has similar effects in Ca^{2+} dynamics to what has been described for vertebrates, this same model could also explain the differences in Ca^{2+} amplitude observed in *Drosophila*.

Finally, in this work, I have assessed the functional equivalence of the vertebrate and *Drosophila* peptides, in light of the extensive evidence suggesting their functional homology. Although the vertebrate peptides do not recapitulate completely the function of the Scl peptides, which is not entirely unexpected given the extent of divergence between these sequences, their effects on the Ca^{2+} dynamics, in either the *scl* mutant rescue, or in over-expression experiments, are in agreement with these peptides having a similar function to that of the Scl peptides. One of the factors, which may contribute to these functional differences, is that in vertebrates it has been shown that although both Sln and Pln are inhibitors of SERCA, the mechanisms of this inhibition are slightly different between these peptides with the inhibitory effect of Pln on SERCA calcium uptake being relieved at high calcium concentrations [103] whereas Sln is inhibitory even at high calcium concentrations [99,107]. As will be discussed in the general discussion, these differences arise from a different structural interaction between Sln and Pln. Since such differences exist already between these paralogues, it is conceivable that more or less subtle differences in the mechanisms of regulation of *SERCA*, or *Ca-P60A* may exist between the vertebrate and *Drosophila* homologues, which may account for the lack of a full functional equivalence, particularly in the case of Sln, which was generally less effective in mimicking the effects of Scl than Pln. Another factor which may be important to consider is that it has been reported in the literature that Sln and Pln can form a ternary complex with SERCA [108], and that this binding of Sln and Pln would be energetically more stable, according to bioinformatics models, than the binding of Sln alone [92]. Since Sln only produces a significant change in arrhythmicity and calcium transient amplitude in an over-expression condition, and therefore in the presence of Scl, a similar mode, in which Sln potentiates the Scl/*Ca-P60A* interaction, may explain this seemingly synergistic effect, if a similar ternary complex occurred between Scl, Sln, and *Ca-P60a*.

In these experiments Pln, was able to recapitulate the effects of Scl on the Ca^{2+} transients, but was unable to completely rescue the arrhythmicity phenotype, while producing a much higher arrhythmicity than Sln and Pln in an over-expression condition. This inability of Pln to fully rescue the arrhythmicity phenotype of *Df scl* mutants could be linked with the much higher arrhythmicity index observed in flies over-expressing Pln. It could be possible, for example, that this peptide, which is relatively different to both Scl and Sln, because of its larger cytoplasmic domain, may have an independent effect on heart rhythmicity in flies, with regards to its regulation of *Ca-P60A*, which would explain the higher arrhythmicity observed during its over-expression, and the lack of a full rescue of the arrhythmic phenotype, despite its behaviour, similar to Scl, regarding the calcium transients in these different conditions. It is important to note however, that the arrhythmicity produced by the over-expression of Pln has a component, which is synergistic with the function of Scl, because the arrhythmicity produced by the expression of Pln is much lower in an *Scl* null background than in a wild-type background (compare Figures 6.8A and 6.8C). In this sense, one needs to bear in mind that the arrhythmicity phenotype, as discussed in the previous chapter, could be considered as a secondary effect of the Ca^{2+} imbalance, whose cause is yet unclear, and therefore the rescue of the Ca^{2+} dynamics should be considered the most relevant result, which in this case, proves the functional homology between these peptides.

3.4 Conclusion

The *scl / pln / sln* family of smORFs, conserved from *Drosophila* to humans, represent an ancient system for the regulation of Ca^{2+} cycling in muscles, and is one of the very few examples of small open reading frame genes conserved across such an evolutionary distance. The Rpl41 ribosomal protein is the only other example of a peptide of under 30 aa, conserved between flies and humans [109,110], although the functional homology between the *Drosophila* Rpl41 and its human orthologue have not been thoroughly assessed yet. These results are therefore important for the field of smORFs, as they show that such conservation is possible in other small peptides, and most importantly, that it is possible to detect it when using the right methods. While this work exposes the limitations of conventional BLAST homology searches when applied to small sequences, it also shows how a more sophisticated homology search method, like

PHYRE2, can be remarkably effective in identifying remote homologues for small sequences, and in unveiling their biological functions.

Chapter VII - General Discussion.

7.1 The Functional homology between Scl Sln and Pln.

The work presented here supports the functional homology between the Scl, Pln, and Sln peptides, with regards to their inhibitory role on Ca^{2+} uptake by SERCA. The regulatory mechanisms leading to this inhibition and its effects in specific muscles and in the whole organism are beginning to be understood in mammals, where the relationship between Pln, Sln and SERCA is the target of extensive research. From my results, it is clear that the regulation of Ca^{2+} by Scl is necessary for the adequate function of the heart in flies. However, the complex patterns of expression, of the different *scl* isoforms show that a regulatory mechanism is already in place to control the expression, and therefore the function of these peptides in different kinds of muscles. My work provides a good starting point for future work to characterise the full extent of the effects of this Scl-mediated Ca^{2+} uptake regulation in flies, and the full extent of homology of this ancestral Ca^{2+} trafficking regulation mechanism between mammals and flies.

7.1a Sln and Pln are reversible inhibitors of SERCA, regulated by their phosphorylation state.

In vertebrates, both Pln and Sln constitute a reversible mechanism to down-regulate the activity of SERCA by lowering its apparent affinity for Ca^{2+} , based on the phosphorylation state of these peptides. This Pln and Sln-dependent inhibition is lifted upon activation of the β -adrenergic pathway by β agonists such as epinephrine [96]. The activation of the β -adrenergic pathway leads to the phosphorylation and inactivation of both Pln and Sln, throughout protein Kinase A (PKA), which mediates the phosphorylation of Pln, at its Ser-16 residue, and throughout the Ca^{2+} /calmodulin dependent kinase II (CaMKII), which mediates the phosphorylation of both Pln and Sln, at their Thr-17 and Thr-5 residues, respectively [111] (Figure 7.1). These

phosphorylation events lead to the inactivation of the inhibitory effects of Sln and Pln on SERCA, and therefore to the upregulation of the SERCA activity, which is reflected by an increase in the force of muscle contraction [96].

The regulation of Scl on Ca-P60A could also be reversible, in which case it would be important to determine whether the regulation of Scl is also dependent on its phosphorylation state. Although the phosphorylation of Pln takes place in its N-terminus region, which is not conserved in Sln or Scl, the Sln threonine-5 residue appears to be conserved in different arthropods (including *Bombus*, *Apis*, *Bombyx*, *Anopheles*, and *Drosophila mojavensis*), but not in *Drosophila melanogaster*. However, a common feature of Scl in arthropods is the presence of serine residues in the N-termini of their sequences; for *Drosophilids*, including *D. melanogaster*, these are in positions 2 or 6. Similarly, threonines 9 and 10 appear to be conserved across insects (Figure 7.2). The presence and conservation of some of these residues in arthropods, which are potential targets of phosphorylation, and the fact that during vertebrate evolution, different residues have been selected for the phosphorylation of Pln and Sln, allow for the idea that Scl could also be regulated by its phosphorylation state. This hypothesis would have to be tested experimentally, by in-vitro phosphorylation assays.

7.1b Regulation of Ca-P60A by a β -adrenergic-like pathway?

Another important element is that *Drosophila* lacks an adrenergic system, with neither epinephrine nor norepinephrine occurring normally in flies. Instead the fly uses octopamine (OA), which performs roughly similar functions in the fly as the adrenergic agonists in mammals, by stimulating different families of adrenergic-like OA receptors [112]. It would therefore be interesting to determine if the OA pathway regulates the inhibition of Ca-P60A by Scl, as the β -adrenergic pathway regulates the inhibition of SERCA by Sln/Pln.

Although OA is found at high levels in the central and peripheral nervous tissues of the fly [113], where it acts as a neurotransmitter, circulating levels of OA have been observed in the haemolymph, particularly during conditions of stress, where OA plays a neurohormonal role [114]. Interestingly, in a recent study aimed at identifying the different effects of OA in the muscle physiology of *Drosophila* larvae, it has been reported that although OA has an effect on synaptic modulation, potentiating neuromuscular transduction, it also has an effect on the contractile force of the muscles

—as measured through a sensitive tension transducer [115]. Importantly, this effect is independent of its neuro-modulatory effect, and is mediated by a molecular mechanism intrinsic to the muscle cell [115]. In that study, it was proposed that the OA-mediated intramuscular effect, leading to muscle contraction strength-enhancement, would most likely occur through the activation of a β -adrenergic-like receptor, localised post-synaptically on the muscle membrane, which would respond by inducing a second messenger system, possibly involving cyclic adenosyl monophosphate (cAMP), and PKA, and which would ultimately act on a yet unidentified target. A hypothesis, which was suggested to explain the intracellular mechanism leading to this increase in contraction strength, implicated the stiffening of the giant muscle protein Titin, which has a structural role in sarcomeres but also in their elastic properties [116]. These observations are interesting, because they show that *Drosophila* muscles can respond directly to OA. One could even go further and hypothesise that Scl may be involved in those intramuscular effects. As shown in this work, *scl* is highly expressed in larval muscles and modulates Ca^{2+} dynamics in hearts. This hypothesis would make sense, considering that such an increase in muscle contraction strength, induced by an intrinsic molecular mechanism within the muscle, and which is activated by the β -adrenergic-like pathway, is in line with the effects of Pln or Sln on mammalian muscles. A similar study, assessing the strength of muscle contraction and calcium dynamics in larval muscles—or cardiac muscles, with a sensitive enough tension transducer—would offer a particularly interesting context in which to study the relation between Scl and the β -adrenergic-like pathway, and therefore the relation between the regulatory pathways that govern the activity of the Scl peptides in flies, and the Pln/Scl peptides in mammals. Furthermore, such a study would provide a unified function for *scl* across different types of muscles.

OA has an effect on *Drosophila* larval muscles, similar to that of adrenaline in mammalian muscles, increasing muscle contractility. My results do not reflect an effect of Scl on the contractility of the heart. It is important to keep in mind, however, that in my case, the contractility was measured by the differences in fractional shortening, between *Scl* mutants and wild-type animals, which may not necessarily reflect the strength with which the cardiomyocytes are contracting. In fact, in mice doubly mutant for Sln and Pln, where the link between Sln, Pln and cardiomyocyte contractility is well established, although young mice present an enhanced Ca^{2+} uptake activity by SERCA,

and higher amplitudes in their Ca^{2+} transients, no differences were observed in the cardiac fractional shortening of those animals. Older mice even presented a lower fractional shortening than wild-type, which is the opposite as would be expected from an enhancement in contractile strength, although this was attributed to the effects of cardiac hypertrophy on these mice [101]. These results suggest that the heart, in *Drosophila* and mammals, may already contract to its maximum capacity in wild-type animals, and therefore it may not be possible to observe larger than wild-type fractional shortenings.

Regarding the behavioural observations presented in this study, showing no differences between *scl* null and wild-type adult flies, and stating that no obvious behavioural issues were observed in larvae, if *Scl* like *Pln* and *Sln*, had an effect on the contractile strength of the muscle, it is not necessarily easy to imagine what such effects would represent for the behaviour of the animal. Interestingly, even though *Sln* and *Pln* are also known to regulate Ca^{2+} in skeletal muscles, *Sln/Pln* double knock down mice are viable, have a normal morphological appearance, and there are no behavioural phenotypes reported for these mutants either, apart from the signs of cardiac hypertrophy in the heart with age [106]. It is possible, however, that a more meticulous analysis of larval motility, and flight capabilities, in *Drosophila*, which would take into account parameters such as crawling or flying speeds, for example, may shed light on the effect of this misregulation of Ca^{2+} dynamics in somatic muscles and on the organism as a whole.

7.1c Mechanistic differences between the *Pln* and *Sln* inhibition of SERCA.

In mammals, *sln* and *pln* are expressed in different kinds of muscles, with *pln* being preferentially expressed in cardiac muscles, and *sln* being preferentially expressed in skeletal muscles and cardiac atrial muscles [117,118]. The different paralogues of *SERCA* themselves are expressed in different tissues, with *SERCA1* being expressed in skeletal muscles, *SERCA2* in both skeletal and cardiac muscles, and *SERCA3* in non muscle cells [97]. The expression of different *SERCA* paralogues has been shown to coincide with the different patterns of *Sln* and *Pln*, suggesting that *Sln* and *Pln* may preferentially inhibit specific *SERCA* paralogues in some contexts [108]. This differential expression of the *pln* and *sln* genes could reflect their different mechanisms regarding the inhibition of *SERCA*. As mentioned before in this work, *Sln* is able to inhibit the apparent Ca^{2+} affinity of *SERCA* at high concentrations, whereas *Pln* cannot. The mechanisms of these differences have recently been addressed by a study which

shows that Sln can bind to SERCA throughout the conformational changes of the ATPase, either in its Ca^{2+} rich form, or in its calcium depleted form [119], even though it does so preferentially in an ATP enriched intermediary state, as the one used to model the Scl and Ca-P60A interaction by F.M.G Pearl. On the other hand Pln is only able to bind the calcium depleted conformation of the ATPase. This implies that Sln can uncouple the ATP pump from Ca^{2+} transport, as it can maintain its inhibition while the ATP is hydrolysed, leading to a “futile” action of the pump, which has been associated with a function of Sln in muscle-based thermogenesis in mammals, and in which Pln is not implicated [120]. This particular mechanism can explain the higher expression of Sln in skeletal muscles, which have a more important contribution to thermogenesis than cardiac muscles. Another important difference is that even though Pln and Sln share extensive sequence similarities in their transmembrane domain and bind to SERCA within the same transmembrane groove, which is in accordance with the computational docking results of F.M.G. Pearl, Sln and Pln have been shown to interact with distinct SERCA residues within that groove [119].

Because the Scl sequence has diverged from those of Sln and Pln, it could be possible that Scl itself interacts with different amino acids within the transmembrane groove of Ca-P60A, for which the Sln peptide may have less affinity. Such a scenario would explain that Sln is still able to bind Ca-P60A, as shown in the co-immunoprecipitations of J.I. Pueyo, but not with the sufficient affinity to entirely fulfil the function of Scl, as shown by the very limited rescue of the *scl* null calcium transients in this work. Pln on the other hand seems to be able to interact with Ca-P60A almost in the same way as Scl, at least with regards to calcium regulation. This result is interesting as it would place the Pln peptide, whose unique function is to dampen the activity of SERCA as opposed to Sln, which also participates in the generation of heat through this inhibition, as more closely related to Scl, despite its longer N-terminal domain. In agreement with these observations, the amino acid alignments for these peptides (Figure 6.3) seem to show consistently more similarity between the arthropod Scl and vertebrate Pln sequences than between Scl and Sln. Altogether, these results could indicate that the ancestral function of these peptides, conserved throughout evolution, was to dampen the activity of SERCA, with Sln then acquiring the ability to co-opt this process to generate heat. On the other hand, endothermy has long been known to exist in insects [121] with the activity of IFMs, where *scl* is expressed, being the main mechanism responsible for the

increase in insect body temperature. This increase in temperature can be dramatic, with the thorax of some flying insects, like bumblebees and moths, being able to reach temperatures above 40°C solely through the activity of IFMs [122]. In most flying insects this high muscle temperature is not only a consequence of IFM activity, but also a requirement to allow the muscles to generate the required wingbeat frequencies necessary for flight [121]. What is particularly interesting is that it has been demonstrated that honeybees have a mechanism to dynamically modulate their metabolic heat production, depending on the temperature of the air [123]. Furthermore, in *Sphinx* moths, it has been shown that the heart plays an essential role in regulating the thoracic temperature, by allowing thermal exchanges between the heat producing thorax and the cooler abdomen of the animals, preventing the thorax from overheating and allowing flight at higher ambient temperatures [122]. Overall, these observations indicate that in some insects, cardiac and IFMs have a functional relationship as, either direct or indirect, thermoregulators. It would be interesting to explore whether *Scl*, as *Sln* in vertebrates, is at all implicated in this thermoregulatory process, as this would provide another evolutionary context for the conservation of this mechanism for the regulation of Ca^{2+} uptake.

In this work I have shown that like *sln* and *pln*, *scl* also presents tissue-specific expression for each of its different isoforms, which may lead to different ratios of ORFA / ORF B peptides being expressed in different muscles (like cardiac muscles and IFMs). The conservation of a mechanism between *Drosophila* and vertebrates that confers such a muscle-specific expression to SERCA regulators is very interesting as it indicates that this mechanism must be ancestral. In fact, if the *Drosophila* peptides were not found to be regulated by their phosphorylation state, nor by the β -adrenergic-like pathway, one could then hypothesise that the ancestral regulation of SERCA activity would have occurred through the regulation of the transcriptional expression of its inhibitors, in a “static” manner, possibly depending on the dosage of inhibitors expressed by different muscles, thereby conferring subtle physiological differences to different muscles. In this case, the acquisition of phosphorylation as a mechanism to regulate the inhibitory effects of the peptides, as it occurs in vertebrates, would have then allowed this regulatory process to become dynamic and responsive to the physiological needs of the organism. In the case of *Drosophila*, and considering the “static” regulation scenario, although my results suggest that both of these peptides are

equivalent in function (because they can both rescue the arrhythmicity and calcium transients of *pncr003;2L* mutants, and induce the same over-expression phenotypes), it is also possible that like Sln and Pln, they may have subtle differences in their mechanism of inhibition of SERCA, which would contribute further to the physiological differentiation of muscles, beyond their mere dosage. It would therefore be interesting to perform similar Ca^{2+} uptake assays on Ca-P60A, with the two Scl peptides, as have been performed in mammals for Pln and Sln [92,94], because such assays, which directly measure the enzymatic activities and rates of the ATPase, may be sensitive enough to detect subtle differences between Scl ORF A and Scl ORB if these exist.

7.1d *Drosophila* as a model for heart disease

Finally, my results show that the misregulation of cardiac calcium in *Drosophila* is linked to heart arrhythmias. This relation between intracellular calcium alterations and heart arrhythmias is not new. In fact, the misregulation of Ca^{2+} represents a hot topic in cardiac research, because it is one of the main implications in heart disease [88], and within this topic, a great focus has been laid on SERCA. As it has been discussed in this work, the SER has the primary function of a Ca^{2+} store, which regulates muscle contraction by releasing Ca^{2+} into the cytosol, and relaxation by active re-uptake of calcium into the SER. SERCA induces muscle relaxation by transporting calcium into the SER, therefore lowering the cytosolic Ca^{2+} concentration. This Ca^{2+} uptake also has the effect of replenishing the SER, with a concentration of Ca^{2+} , which will influence the amounts of Ca^{2+} released in the subsequent contraction events [124]. Other channels like the Plasma membrane Ca^{2+} ATPase (PMCA), or the Sodium /Calcium exchanger channel (NXC) also participate in the maintenance of Ca^{2+} homeostasis in the cell, but their contribution is significantly less than that of SERCA, contributing to 2%, 28%, and 70% of Ca^{2+} depletion from the cytosolic space, respectively [42]. The cytosolic concentrations of Ca^{2+} can not only induce muscle contraction, by acting directly on the sarcomeres, but can also play a feedback role in regulating the currents that lead to muscle contraction in the first place. For example, the L-type voltage-dependent Ca^{2+} channel, which allows the entry of calcium into the cytosol, giving rise to the Ca^{2+} induced Ca^{2+} release by the RyR, is regulated by the cytosolic concentrations of Ca^{2+} [124]. Similarly, Ca^{2+} concentrations have also been shown, in humans, to directly

regulate a specific Na^+ channel (hH1), through the action of CaM, with this interaction leading to a reduction in the inactivation rate of this channel, in similar way as has been associated with heart arrhythmias [87].

On the one hand, cardiac dysfunction is often associated with reduced levels of SERCA expression [125,126,127]. This has led to the research on the effects of gene transfer mediated over-expression of SERCA as a therapeutic alternative for cardiac dysfunctions associated with this down regulation. These kinds of studies have reported promising results, showing the improvement of mice models of cardiac hypertrophy, upon adenoviral gene transfer of SERCA [128]. However, there is a sense of caution for this sort of approach, as over-expression of SERCA has also been shown to significantly increase the risk of acute arrhythmias and death in rat models for myocardial infarction [129].

The *Drosophila* model, as has been presented in this study, should be valuable for future research to elucidate the links between SERCA, Ca^{2+} misregulation and cardiac dysfunction. On the one hand, in *Drosophila*, the misregulation of Ca-P60 has also been shown to be associated with heart arrhythmias [89] and cardiac dysfunction [130]. On the other, I have shown that in the adult *Drosophila* heart, like in vertebrate hearts, the misregulation of Ca^{2+} dynamics, by either the lack or excess of function of *scl*, is associated with a significant increase in cardiac arrhythmias, which can be corrected by either restoring the levels of expression of *scl*, or by modulating those of *Ca-P60A*, as shown in the case of *scl* lack of function. Therefore, in the *Drosophila* adult heart, it is possible to measure consistent Ca^{2+} transient differences between the different genetic conditions for *scl* and the net result of these on heart function. While this system benefits from the genetic tools available for the *Drosophila* model, it is also accessible to pharmacological experimentation, as different compounds such as specific channel inhibitors, can be added to the saline in which the semi-intact preparations bathe, as has previously been reported in larval preparations [131,132]. Furthermore, the work of Jeremy Niven has shown that it is also possible to measure the intracellular action potentials of *Drosophila* cardiomyocytes in these live semi-intact preparations. Altogether, the *Drosophila* heart model represents a truly comprehensive system in which it could be possible to elucidate the effects of SERCA and Ca^{2+} misregulation in the heart by precisely depicting the contribution of the different ion channels and currents, as well as the intracellular molecular pathways governing them.

7.2 Implications of this research on the field of smORFs

7.2a The potential of smORFs

In the general introduction of this thesis, I addressed the challenges that small ORFs represent for genome annotation and portrayed a scenario in which the current repertoire of coding genes for most organisms may be significantly incomplete because we may have missed potentially thousands of small open reading frames, encoding functional peptides such as *tal*, which codes for 11 aa -33 aa peptides, and which have a vital role in *Drosophila*. I addressed how different bioinformatics studies support such a scenario, by predicting the existence of hundreds of putatively coding, and therefore functional smORFs, in organisms that range from yeast to mice, and how further studies have provided experimental evidence that many of these smORFs are functional, leading to morphological defects when over-expressed, in the case of the *Arabidopsis* study [23], or to lack of growth in different mediums when excised, in the case of the yeast study [22].

In this work, I carried out a study, with the aim to determine the specific function of *pncr003;2L*, a gene that was initially annotated as a non-coding RNA, but which had elements suggesting that it may be a protein coding gene, according to one of those bioinformatics studies [19]. Through my work, I have shown that *pncr003;2L* is indeed a protein coding gene, therefore validating the computational prediction methods and I have proven that these peptides, through their evolutionarily conserved mechanism of regulation for Ca^{2+} trafficking during muscle contraction, participate in an ancestral and major cellular process. Overall, these results urge us to carefully consider the potential of smORFs which can be hidden in the genome, and which could have major implications in the biology of the organisms encoding them.

7.2b putatively non-coding RNAs could represent a rich source of smORFs

Recently, the development of techniques which confer the ability to globally assess the transcription of genomes (initially through tilling arrays, and subsequently through RNA sequencing) portray an unexpected scenario where up to 70% of the genome appears to be transcribed in humans and flies [133,134]. A large proportion of that transcriptional signal corresponds to long intergenic non-coding RNAs (lincRNAs), which are putatively non-coding sequences, mostly on the basis of their lack of a long ORF, generally thought to act as “RNA guides” that recruit protein regulatory

complexes to specific genomic loci to control gene expression [135]. Current statistics from the genome reference consortium [136] indicate that in humans, there are 13,564 lincRNAs, a huge number compared with the 20,769 genes annotated as protein coding [137]. In *Drosophila*, the current number of annotated non-coding genes, is also considerable although more modest, with 1,331 lincRNAs and 13,937 coding genes [138,139].

From the example of *scl* presented here, and from that of *tal*, it would be conceivable to believe that such a vast proportion of putatively non-coding genes may represent a rich source of smORFs with important functions. This would in fact seem to be the case, as recent studies based on ribosome profiling (a technique, which uses ribosome protection to nucleases treatment in combination with RNA-seq to obtain a global snapshot of ribosome rich, and therefore translation prone sequences) suggest that at least half of the lincRNAs in mice and zebra fish have translational profiles similar to those of canonical protein-coding genes, which indicates that smORF encoding [140,141]. Therefore, the misannotation of lincRNAs may be widespread, with many smORFs awaiting characterisation in these transcripts. Another important technical development which may influence the discovery of new smORFs, is the development of more sensitive peptidomics techniques, which are able to detect novel small peptides, by using ORF libraries built from RNA-seq readings to analyse the data obtained from mass-spectroscopy, instead of sequences for existing protein databases. A study using these methods has already identified 90 peptides, corresponding to novel smORFs, in a specific human cell line [142], of which 30 are under 30 aa long.

7.2 c Contributions of this work as a case study for the functional characterisation of smORFs

Given that the experimental evidence that supports the existence of large numbers of smORFs seems to be growing, it would be conceivable that, like *scl*, other smORFs may be conserved between remotely-related species. In this sense, this work has shown that standard BLAST searches are rather limited in their ability to identify homologues for sequences of small sizes, and may therefore have to be complemented with the use of a more sophisticated engine such as PHYRE2, which was able to identify the *scl* homologue in humans. The PHYRE2 prediction was particularly convincing after taking into consideration the expression patterns of the *Drosophila* and vertebrate genes, as well as their subcellular localisation. It would therefore be conceivable to use

this kind of information for large scale annotation of smORFs, or functional assays, if this engine was used to search for homology between novel sets of sequences. In that case, in order to support the homology one could compare the RNA-seq expression profiles of putative smORF homologues, in most cases available in data repositories, and their subcellular localisation in cell culture assays. For this to be possible, however, the PHYRE2 engine would have to integrate the possibility to search for similarities between custom libraries of predicted structures corresponding to novel sequences, instead of comparing the predicted structure of a query sequence with the structures and sequences from a library of annotated proteins.

Regarding the use of transposon based strategies to generate mutants for novel smORFs for a reverse genetics approach, as a case study, this work shows that these methods are effective, because it was possible to obtain a null condition for *scl*, but it also exposes some of their weaknesses. Apart from the, apparently remote possibility of having associated alleles (only 0.3-0.5% of the pBac stocks generated for the Exelixis collection were reported to have background mutations [54]), these weaknesses also lay on the extensive work required to generate and map the new mutations, and importantly, on the fact that other genes can also be affected by them, such as the *CG31739*, *CG13282* and *CG1328* genes in this case. These kinds of issues may justify the lack of functional studies undertaken on novel genes, such as smORF genes, despite having transposon insertions in, or near to, most genes in the *Drosophila* genome. As new methods and tools to target specific genes are emerging in the field of genetics, such as the recent improvement and simplification of the ends-out homologous recombination method, which used to be extremely laborious and very ineffective (requiring several generations of genetic crosses, and the screening hundreds of thousands of flies, in order to obtain a positive recombinant [143]), but would now yield a fair amount of positive recombinants in only a couple of generations of flies [144], it is possible that the functional characterisation of novel genes, and particularly novel smORF genes, may become a more feasible, and wide-spread practice in the near future.

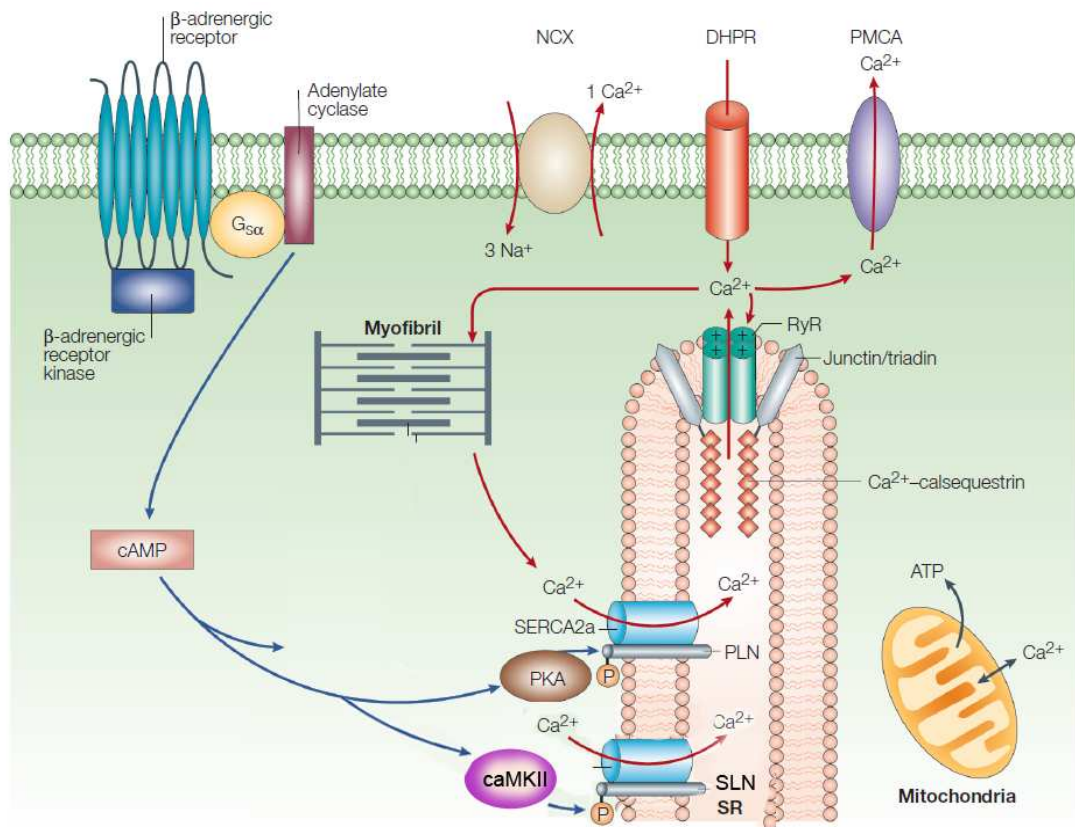


Figure 7.1

Figure 7.1: The inhibition of SERCA2a in cardiac muscles by the Sln and Pln peptides is regulated by the β -adrenergic in vertebrates. Diagram representing the regulation of Sln and Pln by the β -adrenergic pathway, modified from [96]. Sln and Pln are inhibitors of SERCA2a, in cardiac muscles. The signal from the β -adrenergic receptor proceeds through G_s proteins to stimulate the formation of cyclic AMP (cAMP) by adenyl cyclase. Elevations of cAMP activate cAMP dependent protein kinase A (PKA) and Ca^{2+} calmodulin dependent kinase II (CaMKII). PKA and CaMKII phosphorylate Phospholamban (PLN) and Sarcolipin and (SLN), and release their inhibitory effect on SERCA2a, leading to increased Ca^{2+} uptake, and release from the cytoplasmic stores, and to increased cardiomyocyte contractility.

| | | |
|-------------------|--|----|
| Onychiurus | -----M-----EPKGSSEYNLVVITYGILILLGLVWLLYQGFG----- | 33 |
| Pissodes | -----MAG-----AHSEGKSLITTYLILIALLLVLWALYAGMF----- | 33 |
| Sitophilus | -----MPP-----HASEGKSLITTYLILIALLLVLWGLYAGMF----- | 33 |
| D. melanogaster_B | -----MN-----EAKSLFTTFLILAFLLFLLYAFYEAAF----- | 29 |
| D. mojavensis_B | -----MN-----EAKSLFTTFLILAFLLFLLYAFYEAAF----- | 29 |
| Sarcophaga_B | -----MN-----EAKSLFTTFVILVFLLSLLYIFYEIAF----- | 29 |
| Sarcophaga_A | -----MS-----EGKSLSTFLILAVLLSLLYILYAIF----- | 28 |
| Anopheles_A | -----MS-----ETKNLMTTFILIFLWLLLYLVYEGFYQPEM----- | 33 |
| D. melanogaster_A | -----MS-----EARNLFTTFGILAILLFFLYLIYAVL----- | 28 |
| D. mojavensis_A | -----MS-----EATNLFTTFGILAILLFFLWIIYAVL----- | 28 |
| Bombyx_A | -----MP-----NHETTNIATVCILVLLASLWLLYSGMF----- | 31 |
| Bombyx_B | -----MPI-----NAETVNLAATYMLVILLFFLWLLYSTVF----- | 32 |
| Bombyx_C | -----MAL-----ASEGTNLVFTYFILLLLVSLWMLYS-AF----- | 31 |
| Trichoplusia | -----MN-----SAEGTNLVATYCILLLLVSLWLVYS-AF----- | 30 |
| Bombus | -----MPQA-----AHETKNILTTYFILILLICLWLLYSGLF----- | 33 |
| Apis | -----MPQA-----AHETKNILTTYFILILLICLWLLYSGLFV----- | 34 |
| Nasonia | -----MTGTAAQKRSSLKIPQ-----SHESKNILTTYFILLLLISLYLLYSGIY----- | 47 |
| Triops | -----MDQVTALENREAKSLVVNVLVIILLGILWLLYTGA----- | 36 |
| Daphnia_pulex | -----MNNP-----EHAHAKSLIINYVVIILLSLWLLCEGM----- | 33 |
| Daphnia_Magna | -----MNNP-----EHAHAKSLIINYVVIILLSLWLLYEGM----- | 33 |
| Ixodes_A | -----MDT-----KEGQSLFINYLILIVLLILWVLYVGVDK----- | 32 |
| Ixodes_B | -----MD-----EASSLVINYGVLIILLLILWVLYTGM----- | 29 |
| Limulus | -----MIDP-----EVQSLVMNYVILIVLLVILWILYTG----- | 30 |
| Danio_sln | -----MER-----STQELFLNFMIVLITVLLMWLLVKQYQYD----- | 32 |
| Danio_pln | MEKVQYMTAAAIRRASTMEVPQQAQ--NM-QELFVNFCILILCLLIYIIVLLIS----- | 52 |
| Human_sln | -----MGI-----NTRELFLNFTIVLITVILMWLLVRSYQY----- | 31 |
| Human_pln | MEKVQYLTRSAIRRASTIEMPQQAQ--KL-QNLFINFCILILCLLLICLIIVMLL----- | 52 |

Figure 7.2

Figure 7.2: Possible conserved phosphorylation sites in the Scl family of peptides: Alignment highlighting the N-terminus Threonine and Serine residues (bold and underlined) known to be phosphorylated in vertebrates, and showing patterns of conservation in arthropods, suggesting that the arthropod peptides may be subjected to the same phosphorylation-dependent regulation as their vertebrate homologues.

References

1. Johan Henrik W (1975) An analysis of Wilhelm Johannsen's genetical genotype "term" 1909–26. *Hereditas* 79: 1-4.
2. WINGE Ö (1958) WILHELM JOHANNSEN: The Creator of the Terms Gene, Genotype, Phenotype and Pure Line. *Journal of Heredity* 49: 83-88.
3. Allen JE, Pertea M, Salzberg SL (2004) Computational gene prediction using multiple sources of evidence. *Genome Research* 14: 142-148.
4. Basrai MA, Hieter P, Boeke JD (1997) Small Open Reading Frames: Beautiful Needles in the Haystack. *Genome Research* 7: 768-771.
5. Sharp PM, Li WH (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15: 1281-1295.
6. Claverie JM, Poirot O, Lopez F (1997) The difficulty of identifying genes in anonymous vertebrate sequences. *Comput Chem* 21: 203-214.
7. Fickett JW (1995) ORFs and genes: how strong a connection? *J Comput Biol* 2: 117-123.
8. Brent MR (2005) Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res* 15: 1777-1786.
9. Dujon B, Alexandraki D, Andre B, Ansorge W, Baladron V, et al. (1994) Complete DNA sequence of yeast chromosome XI. *Nature* 369: 371-378.
10. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. *Science* 309: 1559-1563.
11. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP (2007) Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *Plos Biology* 5: 1052-1062.
12. Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, et al. (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biology* 9: 660-U687.
13. Pueyo JI, Couso JP (2008) The 11-aminoacid long Tarsal-less peptides trigger a cell signal in *Drosophila* leg development. *Developmental Biology* 324: 192-201.
14. Pueyo JI, Couso JP (2011) Tarsal-less peptides control Notch signalling through the Shavenbaby transcription factor. *Dev Biol* 355: 183-193.
15. Artavanis-Tsakonas S, Rand MD, Lake RJ (1999) Notch signaling: cell fate control and signal integration in development. *Science* 284: 770-776.
16. Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, et al. (2010) Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* 329: 336-339.
17. Hanada K, Zhang X, Borevitz JO, Li WH, Shiu SH (2007) A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Research* 17: 632-640.
18. Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, et al. (2006) The abundance of short proteins in the mammalian proteome. *Plos Genetics* 2: 515-528.

19. Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP (2011) Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol* 12: R118.
20. Kessler MM, Zeng Q, Hogan S, Cook R, Morales AJ, et al. (2003) Systematic Discovery of New Genes in the *Saccharomyces cerevisiae* Genome. *Genome Res* 13: 264-271.
21. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE (2008) Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol* 70: 1487-1501.
22. Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au W-C, et al. (2006) Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res* 16: 365-373.
23. Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, et al. (2013) Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci U S A* 110: 2395-2400.
24. Tupy JL, Bailey AM, Dailey G, Evans-Holm M, Siebel CW, et al. (2005) Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America* 102: 5495-5500.
25. Pueyo JI, Couso JP (2008) The 11-aminoacid long Tarsal-less peptides trigger a cell signal in *Drosophila* leg development. *Dev Biol* 324: 192-201.
26. Fink M, Callol-Massot C, Chu A, Ruiz-Lozano P, Izpisua Belmonte JC, et al. (2009) A new method for detection and quantification of heartbeat parameters in *Drosophila*, zebrafish, and embryonic mouse hearts. *Biotechniques* 46: 101-113.
27. Ocorr K, Reeves NL, Wessells RJ, Fink M, Chen HS, et al. (2007) KCNQ potassium channel mutations cause cardiac arrhythmias in *Drosophila* that mimic the effects of aging. *Proc Natl Acad Sci U S A* 104: 3943-3948.
28. Ocorr KA, Crawley T, Gibson G, Bodmer R (2007) Genetic variation for cardiac dysfunction in *Drosophila*. *PLoS One* 2: e601.
29. Vogler G, Ocorr K (2009) Visualizing the beating heart in *Drosophila*. *J Vis Exp*.
30. Kelley LA, Sternberg MJ (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* 4: 363-371.
31. Lewis EB (1960) A new standard food medium. *Drosophila Information Service* 34.
32. Kondo T, Inagaki S, Yasuda K, Kageyama Y (2006) Rapid construction of *Drosophila* RNAi transgenes using pRISE, a P-element-mediated transformation vector exploiting an in vitro recombination system. *Genes Genet Syst* 81: 129-134.
33. Magny EG, Pueyo JI, Pearl FM, Cespedes MA, Niven JE, et al. (2013) Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. *Science*.
34. Lin N, Badie N, Yu L, Abraham D, Cheng H, et al. (2011) A method to measure myocardial calcium handling in adult *Drosophila*. *Circ Res* 108: 1306-1315.
35. Sanyal S, Consoulas C, Kuromi H, Basole A, Mukai L, et al. (2005) Analysis of conditional paralytic mutants in *Drosophila* sarco-endoplasmic reticulum calcium ATPase reveals novel mechanisms for regulating membrane excitability. *Genetics* 169: 737-750.
36. Cavener DR (1987) Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res* 15: 1353-1361.
37. Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* 361: 13-37.

38. Bate M (1990) The embryonic development of larval muscles in *Drosophila*. *Development* 110: 791-804.
39. Fischer JA, Giniger E, Maniatis T, Ptashne M (1988) GAL4 activates transcription in *Drosophila*. *Nature* 332: 853-856.
40. Bateman JR, Lee AM, Wu CT (2006) Site-specific transformation of *Drosophila* via phiC31 integrase-mediated cassette exchange. *Genetics* 173: 769-777.
41. Razzaq A, Robinson IM, McMahon HT, Skepper JN, Su Y, et al. (2001) Amphiphysin is necessary for organization of the excitation-contraction coupling machinery of muscles, but not for synaptic vesicle endocytosis in *Drosophila*. *Genes Dev* 15: 2967-2979.
42. Gwathmey JK, Yerevanian AI, Hajjar RJ (2011) Cardiac gene therapy with SERCA2a: from bench to bedside. *J Mol Cell Cardiol* 50: 803-812.
43. Pi H, Lee LW, Lo SJ (2009) New insights into polycistronic transcripts in eukaryotes. *Chang Gung Med J* 32: 494-498.
44. Blumenthal T (2004) Operons in eukaryotes. *Brief Funct Genomic Proteomic* 3: 199-211.
45. Fire A, Xu SQ, Montgomery MK, Kostas SA, Driver SE, et al. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391: 806-811.
46. Perrimon N, Mathey-Prevot B (2007) Applications of high-throughput RNA interference screens to problems in cell and developmental biology. *Genetics* 175: 7-16.
47. (2003) Whither RNAi? *Nat Cell Biol* 5: 489-490.
48. Sarov M, Stewart AF (2005) The best control for the specificity of RNAi. *Trends Biotechnol* 23: 446-448.
49. Fedoroff NV (2012) Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* 338: 758-767.
50. Cooley L, Berg C, Kelley R, McKearin D, Spradling A (1989) Identifying and cloning *Drosophila* genes by single P element insertional mutagenesis. *Prog Nucleic Acid Res Mol Biol* 36: 99-109.
51. Cooley L, Berg C, Spradling A (1988) Controlling P element insertional mutagenesis. *Trends Genet* 4: 254-258.
52. Cooley L, Kelley R, Spradling A (1988) Insertional mutagenesis of the *Drosophila* genome with single P elements. *Science* 239: 1121-1128.
53. Bellen HJ, Levis RW, He YC, Carlson JW, Evans-Holm M, et al. (2011) The *Drosophila* Gene Disruption Project: Progress Using Transposons With Distinctive Site Specificities. *Genetics* 188: 731-U341.
54. Thibault ST, Singer MA, Miyazaki WY, Milash B, Dompe NA, et al. (2004) A complementary transposon tool kit for *Drosophila melanogaster* using P and piggyBac. *Nature Genetics* 36: 283-287.
55. Cox MM (1988) FLP Site-Specific Recombination System of *Saccharomyces cerevisiae*. *Genetic Recombination*. Washington D.C.: American Society for Microbiology.
56. Parks AL, Cook KR, Belvin M, Dompe NA, Fawcett R, et al. (2004) Systematic generation of high-resolution deletion coverage of the *Drosophila melanogaster* genome. *Nature Genetics* 36: 288-292.
57. Cripps RM, Ball E, Stark M, Lawn A, Sparrow JC (1994) Recovery of dominant, autosomal flightless mutants of *Drosophila melanogaster* and identification of a new gene required for normal muscle structure and function. *Genetics* 137: 151-164.

58. Nongthomba U, Cummins M, Clark S, Vigoreaux JO, Sparrow JC (2003) Suppression of muscle hypercontraction by mutations in the myosin heavy chain gene of *Drosophila melanogaster*. *Genetics* 164: 209-222.
59. Kronert WA, Acebes A, Ferrus A, Bernstein SI (1999) Specific myosin heavy chain mutations suppress troponin I defects in *Drosophila* muscles. *J Cell Biol* 144: 989-1000.
60. Nongthomba U, Clark S, Cummins M, Ansari M, Stark M, et al. (2004) Troponin I is required for myofibrillogenesis and sarcomere formation in *Drosophila* flight muscle. *J Cell Sci* 117: 1795-1805.
61. Landis G, Tower J (1999) The *Drosophila* chiffon gene is required for chorion gene amplification, and is related to the yeast Dbf4 regulator of DNA replication and cell cycle. *Development* 126: 4281-4293.
62. Hastings GA, Emerson CP, Jr. (1991) Myosin functional domains encoded by alternative exons are expressed in specific thoracic muscles of *Drosophila*. *J Cell Biol* 114: 263-276.
63. Chen M, Manley JL (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* 10: 741-754.
64. Kabat JL, Barberan-Soler S, McKenna P, Clawson H, Farrer T, et al. (2006) Intronic alternative splicing regulators identified by comparative genomics in nematodes. *PLoS Comput Biol* 2: e86.
65. Sawicka K, Bushell M, Spriggs KA, Willis AE (2008) Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem Soc Trans* 36: 641-647.
66. Weiss A, Leinwand LA (1996) The mammalian myosin heavy chain gene family. *Annu Rev Cell Dev Biol* 12: 417-439.
67. Barany M (1967) ATPase activity of myosin correlated with speed of muscle shortening. *J Gen Physiol* 50: Suppl:197-218.
68. Kronert WA, Edwards KA, Roche ES, Wells L, Bernstein SI (1991) Muscle-specific accumulation of *Drosophila* myosin heavy chains: a splicing mutation in an alternative exon results in an isoform substitution. *Embo J* 10: 2479-2488.
69. Rubin GM (1990) *Drosophila* - a Laboratory Handbook - Ashburner, M. *Nature* 348: 366-366.
70. Koana T, Hotta Y (1978) Isolation and characterization of flightless mutants in *Drosophila melanogaster*. *J Embryol Exp Morphol* 45: 123-143.
71. Deak, II (1977) Mutations of *Drosophila melanogaster* that affect muscles. *J Embryol Exp Morphol* 40: 35-63.
72. Nongthomba U, Ramachandra NB (1999) A direct screen identifies new flight muscle mutants on the *Drosophila* second chromosome. *Genetics* 153: 261-274.
73. Chen B, Chu T, Harms E, Gergen JP, Strickland S (1998) Mapping of *Drosophila* mutations using site-specific male recombination. *Genetics* 149: 157-163.
74. Bodmer R (1993) The gene tinman is required for specification of the heart and visceral muscles in *Drosophila*. *Development* 118: 719-729.
75. Lints TJ, Parsons LM, Hartley L, Lyons I, Harvey RP (1993) Nkx-2.5: a novel murine homeobox gene expressed in early heart progenitor cells and their myogenic descendants. *Development* 119: 419-431.
76. Cripps RM, Olson EN (2002) Control of cardiac development by an evolutionarily conserved transcriptional network. *Dev Biol* 246: 14-28.
77. Jentsch TJ (2000) Neuronal KCNQ potassium channels: physiology and role in disease. *Nat Rev Neurosci* 1: 21-30.

78. Wolf MJ, Rockman HA (2011) *Drosophila*, genetic screens, and cardiac function. *Circ Res* 109: 794-806.
79. Rizki TM (1978) The circulatory system and associated cells and tissues. In: Wright MAaTRF, editor. *The Genetics and Biology of Drosophila*. New-York: Academic Press. pp. pp. 397-452.
80. Dulcis D, Levine RB (2003) Innervation of the heart of the adult fruit fly, *Drosophila melanogaster*. *J Comp Neurol* 465: 560-578.
81. Dulcis D, Levine RB (2005) Glutamatergic innervation of the heart initiates retrograde contractions in adult *Drosophila melanogaster*. *J Neurosci* 25: 271-280.
82. Lo PC, Frasch M (2001) A role for the COUP-TF-related gene seven-up in the diversification of cardioblast identities in the dorsal vessel of *Drosophila*. *Mech Dev* 104: 49-60.
83. Lehmacher C, Abeln B, Paululat A (2012) The ultrastructure of *Drosophila* heart cells. *Arthropod Struct Dev* 41: 459-474.
84. Tian L, Hires SA, Mao T, Huber D, Chiappe ME, et al. (2009) Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nat Methods* 6: 875-881.
85. Akerboom J, Rivera JD, Guilbe MM, Malave EC, Hernandez HH, et al. (2009) Crystal structures of the GCaMP calcium sensor reveal the mechanism of fluorescence signal change and aid rational design. *J Biol Chem* 284: 6455-6464.
86. Bers DM (2002) Calcium and cardiac rhythms: physiological and pathophysiological. *Circ Res* 90: 14-17.
87. Tan HL, Kupersmidt S, Zhang R, Stepanovic S, Roden DM, et al. (2002) A calcium sensor in the sodium channel modulates cardiac excitability. *Nature* 415: 442-447.
88. Clusin WT (2003) Calcium and Cardiac Arrhythmias: DADs, EADs, and Alternans. *Critical Reviews in Clinical Laboratory Sciences* 40: 337-375.
89. Sanyal S, Jennings T, Dowse H, Ramaswami M (2006) Conditional mutations in SERCA, the Sarco-endoplasmic reticulum Ca²⁺-ATPase, alter heart rate and rhythmicity in *Drosophila*. *J Comp Physiol B* 176: 253-263.
90. Ohlson T, Wallner B, Elofsson A (2004) Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins* 57: 188-197.
91. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389-3402.
92. Asahi M, Sugita Y, Kurzydowski K, De Leon S, Tada M, et al. (2003) Sarcolipin regulates sarco(endo)plasmic reticulum Ca²⁺-ATPase (SERCA) by binding to transmembrane helices alone or in association with phospholamban. *Proc Natl Acad Sci U S A* 100: 5040-5045.
93. Babu GJ, Bhupathy P, Petrashevskaya NN, Ziolo MT, Periasamy M (2006) Sarcolipin regulates atrial calcium transport and contractility. *Circulation* 114: 232-232.
94. Bhupathy P, Babu GJ, Periasamy M (2007) Sarcolipin and phospholamban as regulators of cardiac sarcoplasmic reticulum Ca²⁺ ATPase. *J Mol Cell Cardiol* 42: 903-911.
95. Babu GJ, Bhupathy P, Petrashevskaya NN, Wang HL, Raman S, et al. (2006) Targeted overexpression of sarcolipin in the mouse heart decreases sarcoplasmic

- reticulum calcium transport and cardiac contractility. *Journal of Biological Chemistry* 281: 3972-3979.
96. MacLennan DH, Kranias EG (2003) Phospholamban: a crucial regulator of cardiac contractility. *Nat Rev Mol Cell Biol* 4: 566-577.
 97. Periasamy M, Kalyanasundaram A (2007) SERCA pump isoforms: their role in calcium transport and disease. *Muscle Nerve* 35: 430-442.
 98. Vazquez-Martinez O, Canedo-Merino R, Diaz-Munoz M, Riesgo-Escovar JR (2003) Biochemical characterization, distribution and phylogenetic analysis of *Drosophila melanogaster* ryanodine and IP3 receptors, and thapsigargin-sensitive Ca^{2+} ATPase. *J Cell Sci* 116: 2483-2494.
 99. Asahi M, Otsu K, Nakayama H, Hikoso S, Takeda T, et al. (2004) Cardiac-specific overexpression of sarcolipin inhibits sarco(endo)plasmic reticulum Ca^{2+} ATPase (SERCA2a) activity and impairs cardiac function in mice. *Proc Natl Acad Sci U S A* 101: 9199-9204.
 100. Babu GJ, Bhupathy P, Timofeyev V, Petrashevskaya NN, Reiser PJ, et al. (2007) Ablation of sarcolipin enhances sarcoplasmic reticulum calcium transport and atrial contractility. *Proc Natl Acad Sci U S A* 104: 17867-17872.
 101. Shanmugam M, Gao S, Hong C, Fefelova N, Nowycky MC, et al. (2011) Ablation of phospholamban and sarcolipin results in cardiac hypertrophy and decreased cardiac contractility. *Cardiovasc Res* 89: 353-361.
 102. Winther AM, Bublitz M, Karlsen JL, Moller JV, Hansen JB, et al. (2013) The sarcolipin-bound calcium pump stabilizes calcium sites exposed to the cytoplasm. *Nature* 495: 265-269.
 103. Kadambi VJ, Ponniah S, Harrer JM, Hoit BD, Dorn GW, 2nd, et al. (1996) Cardiac-specific overexpression of phospholamban alters calcium kinetics and resultant cardiomyocyte mechanics in transgenic mice. *J Clin Invest* 97: 533-539.
 104. Gyorke S, Terentyev D (2008) Modulation of ryanodine receptor by luminal calcium and accessory proteins in health and cardiac disease. *Cardiovasc Res* 77: 245-255.
 105. Herrmann-Frank A, Lehmann-Horn F (1996) Regulation of the purified Ca^{2+} release channel/ryanodine receptor complex of skeletal muscle sarcoplasmic reticulum by luminal calcium. *Pflugers Arch* 432: 155-157.
 106. Shanmugam M, Gao S, Hong C, Fefelova N, Nowycky MC, et al. (2010) Ablation of phospholamban and sarcolipin results in cardiac hypertrophy and decreased cardiac contractility. *Cardiovasc Res* 89: 353-361.
 107. Babu GJ, Bhupathy P, Petrashevskaya NN, Wang H, Raman S, et al. (2006) Targeted overexpression of sarcolipin in the mouse heart decreases sarcoplasmic reticulum calcium transport and cardiac contractility. *J Biol Chem* 281: 3972-3979.
 108. Fajardo VA, Bombardier E, Vigna C, Devji T, Bloemberg D, et al. (2013) Co-Expression of SERCA Isoforms, Phospholamban and Sarcolipin in Human Skeletal Muscle Fibers. *PLoS One* 8: e84304.
 109. Klinge S, Voigts-Hoffmann F, Leibundgut M, Ban N (2012) Atomic structures of the eukaryotic ribosome. *Trends Biochem Sci* 37: 189-198.
 110. Rugjee KN, Roy Chaudhury S, Al-Jubran K, Ramanathan P, Matina T, et al. (2013) Fluorescent protein tagging confirms the presence of ribosomal proteins at *Drosophila* polytene chromosomes. *PeerJ* 1: e15.

111. Traaseth NJ, Ha KN, Verardi R, Shi L, Buffy JJ, et al. (2008) Structural and dynamic basis of phospholamban and sarcolipin inhibition of Ca(2+)-ATPase. *Biochemistry* 47: 3-13.
112. Evans PD, Maqueira B (2005) Insect octopamine receptors: a new classification scheme based on studies of cloned *Drosophila* G-protein coupled receptors. *Invert Neurosci* 5: 111-118.
113. Roeder T (1999) Octopamine in invertebrates. *Prog Neurobiol* 59: 533-561.
114. Farooqui T (2007) Octopamine-mediated neuromodulation of insect senses. *Neurochem Res* 32: 1511-1529.
115. Ormerod KG, Hadden JK, Deady LD, Mercier AJ, Krans JL (2013) Action of octopamine and tyramine on muscles of *Drosophila melanogaster* larvae. *J Neurophysiol* 110: 1984-1996.
116. Labeit S, Kolmerer B (1995) Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science* 270: 293-296.
117. Babu GJ, Bhupathy P, Carnes CA, Billman GE, Periasamy M (2007) Differential expression of sarcolipin protein during muscle development and cardiac pathophysiology. *J Mol Cell Cardiol* 43: 215-222.
118. Vangheluwe P, Schuermans M, Zador E, Waelkens E, Raeymaekers L, et al. (2005) Sarcolipin and phospholamban mRNA and protein expression in cardiac and skeletal muscle of different species. *Biochem J* 389: 151-159.
119. Sahoo SK, Shaikh SA, Sopariwala DH, Bal NC, Periasamy M (2013) Sarcolipin protein interaction with sarco(endo)plasmic reticulum Ca²⁺ ATPase (SERCA) is distinct from phospholamban protein, and only sarcolipin can promote uncoupling of the SERCA pump. *J Biol Chem* 288: 6881-6889.
120. Bal NC, Maurya SK, Sopariwala DH, Sahoo SK, Gupta SC, et al. (2012) Sarcolipin is a newly identified regulator of muscle-based thermogenesis in mammals. *Nat Med* 18: 1575-1579.
121. Heinrich B (1974) Thermoregulation in endothermic insects. *Science* 185: 747-756.
122. Heinrich B (1970) Thoracic Temperature Stabilization by Blood Circulation in a Free-Flying Moth. *Science* 168: 580-582.
123. Harrison JF, Fewell JH, Roberts SP, Hall HG (1996) Achievement of thermal stability by varying metabolic heat production in flying honeybees. *Science* 274: 88-90.
124. Eisner DA, Choi HS, Diaz ME, O'Neill SC, Trafford AW (2000) Integrative analysis of calcium cycling in cardiac muscle. *Circ Res* 87: 1087-1094.
125. Arai M, Matsui H, Periasamy M (1994) Sarcoplasmic reticulum gene expression in cardiac hypertrophy and heart failure. *Circ Res* 74: 555-564.
126. Arai M, Alpert NR, MacLennan DH, Barton P, Periasamy M (1993) Alterations in sarcoplasmic reticulum gene expression in human heart failure. A possible mechanism for alterations in systolic and diastolic properties of the failing myocardium. *Circ Res* 72: 463-469.
127. Schmidt U, Hajjar RJ, Helm PA, Kim CS, Doye AA, et al. (1998) Contribution of abnormal sarcoplasmic reticulum ATPase activity to systolic and diastolic dysfunction in human heart failure. *J Mol Cell Cardiol* 30: 1929-1937.
128. Hajjar RJ, Kang JX, Gwathmey JK, Rosenzweig A (1997) Physiological effects of adenoviral gene transfer of sarcoplasmic reticulum calcium ATPase in isolated rat myocytes. *Circulation* 95: 423-429.
129. Chen Y, Escoubet B, Prunier F, Amour J, Simonides WS, et al. (2004) Constitutive cardiac overexpression of sarcoplasmic/endoplasmic reticulum Ca²⁺-ATPase

- delays myocardial failure after myocardial infarction in rats at a cost of increased acute arrhythmias. *Circulation* 109: 1898-1903.
130. Abraham DM, Wolf MJ (2013) Disruption of sarcoendoplasmic reticulum calcium ATPase function in *Drosophila* leads to cardiac dysfunction. *PLoS One* 8: e77785.
 131. Cooper AS, Rymond KE, Ward MA, Bocook EL, Cooper RL (2009) Monitoring heart function in larval *Drosophila melanogaster* for physiological studies. *J Vis Exp*.
 132. Johnson E, Sherry T, Ringo J, Dowse H (2002) Modulation of the cardiac pacemaker of *Drosophila*: cellular mechanisms. *J Comp Physiol B* 172: 227-236.
 133. Willingham AT, Dike S, Cheng J, Manak JR, Bell I, et al. (2006) Transcriptional landscape of the human and fly genomes: nonlinear and multifunctional modular model of transcriptomes. *Cold Spring Harb Symp Quant Biol* 71: 101-110.
 134. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. (2012) Landscape of transcription in human cells. *Nature* 489: 101-108.
 135. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 22: 1775-1789.
 136. Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931-945.
 137. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 22: 1760-1774.
 138. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203-218.
 139. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287.
 140. Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of Mammalian proteomes. *Cell* 147: 789-802.
 141. Chew GL, Pauli A, Rinn JL, Regev A, Schier AF, et al. (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* 140: 2828-2834.
 142. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, et al. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* 9: 59-64.
 143. Huang J, Zhou W, Dong W, Watson AM, Hong Y (2009) From the Cover: Directed, efficient, and versatile modifications of the *Drosophila* genome by genomic engineering. *Proc Natl Acad Sci U S A* 106: 8284-8289.
 144. Baena-Lopez LA, Alexandre C, Mitchell A, Pasakarnis L, Vincent JP (2013) Accelerated homologous recombination and subsequent genome modification in *Drosophila*. *Development* 140: 4818-4825.

Annexes

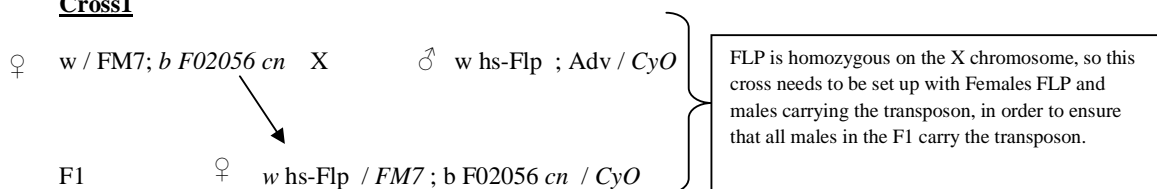
| Currently annotated genes coding for peptides under 30 aa long in <i>Drosophila melanogaster</i> | | |
|---|---------|---|
| size(aa) | name | function |
| 11 | tal-1A | actin filament organisation; morphogenesis of an epithelium; imaginal disc-derived wing morphogenesis |
| 11 | tal-2A | |
| 11 | tal-3A | |
| 21 | CG43178 | unknown |
| 21 | CG43200 | unknown |
| 22 | CG43172 | unknown |
| 23 | CG43171 | unknown |
| 23 | CG43201 | unknown |
| 25 | RpL41 | ribosomal protein |
| 25 | Acp98AB | accessory gland protein |

Annex 1

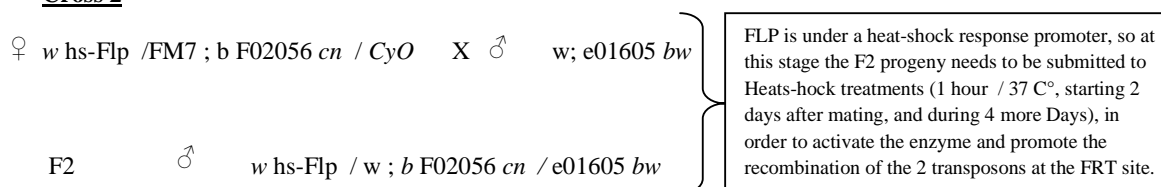
Annex 1: List of currently annotated genes coding for peptides under 30 amino acids long in *Drosophila melanogaster*.

Generation of an FRT-mediated specific deficiency:

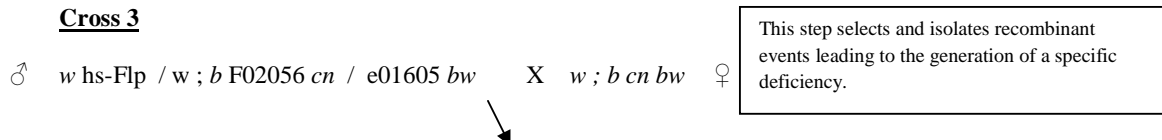
Cross1



Cross 2



Cross 3

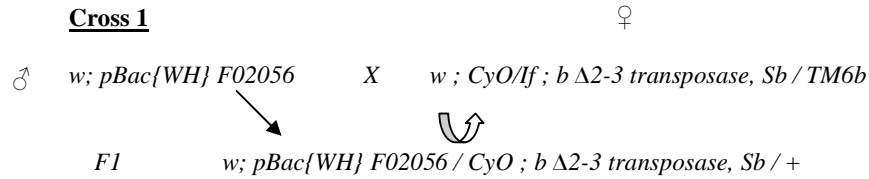


4 possible classes can be obtained:

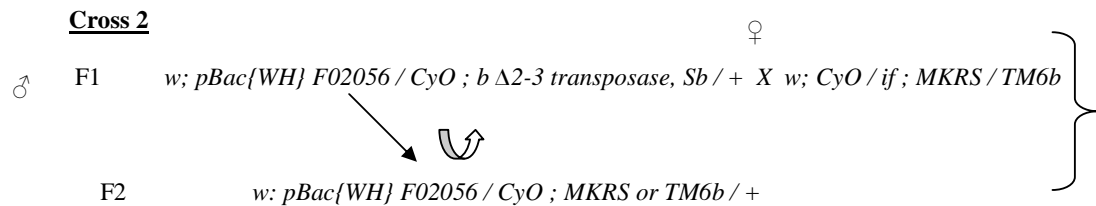
| chromosome class: | genotype: | genetic markers: | phenotype: |
|--|------------------------------------|------------------------|------------------------------|
| parental | $w ; b F02056 cn / b cn bw$ | <i>black, cinnabar</i> | black cuticle, orange eye |
| parental | $w ; e01605 bw / b cn bw$ | <i>brown</i> | pale pink eye |
| recombinant: duplication by-product | $w ; b F02056-e01605 bw / b cn bw$ | <i>black, brown</i> | black cuticle, pale pink eye |
| recombinant: specific deficiency | $w ; e01605-F02056 cn / b cn bw$ | <i>cinnabar</i> | dark orange eye |

Make stable stocks, over *CyO* Balancer, confirm by PCR

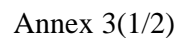
Reversion protocol using a Pbac F02056 -element and the pbac transposase :

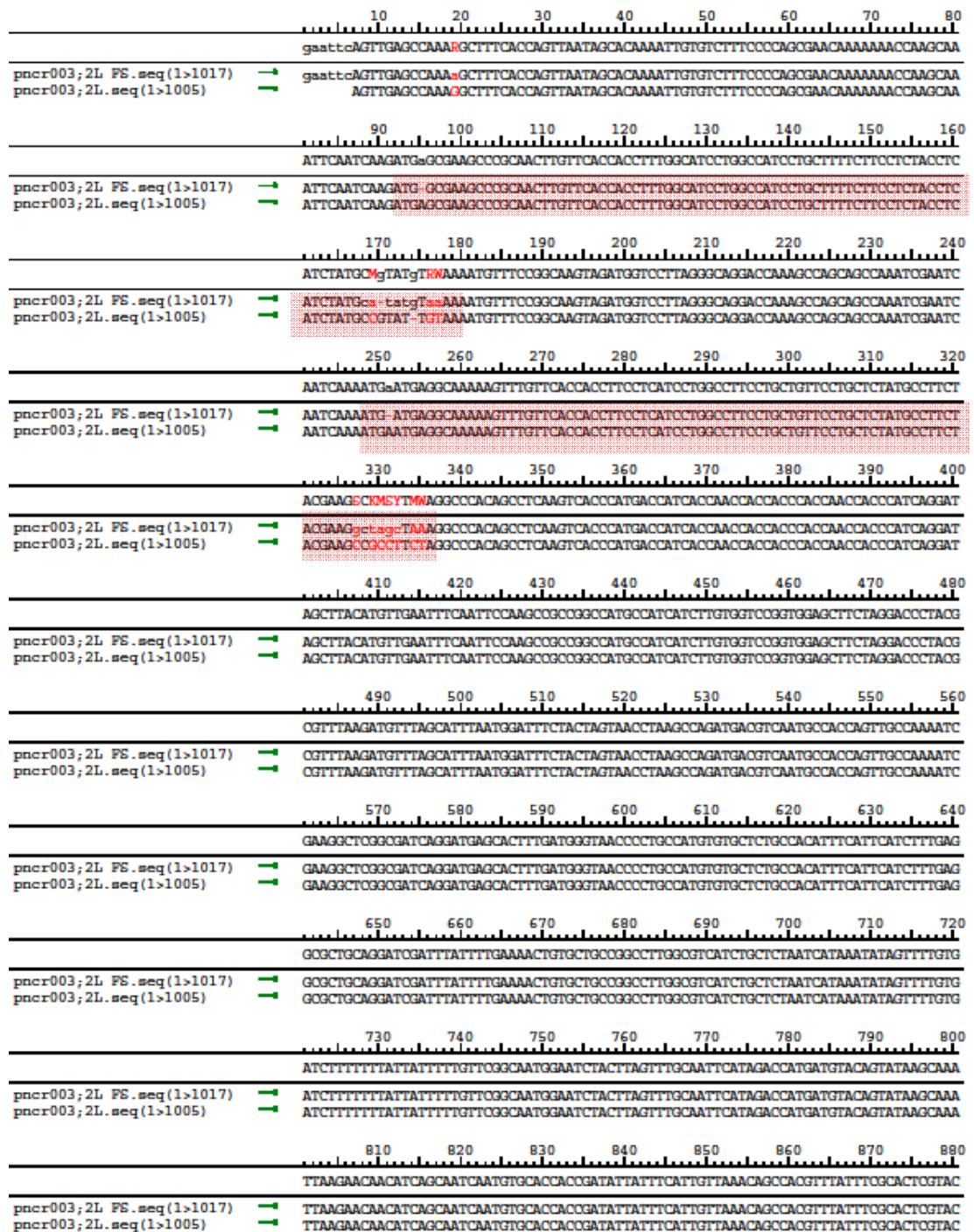


The F1 is mosaic as the transposition event doesn't occur in all the cells of the flies. A reversion will only be maintained in the progeny if it occurred in a germ-cell, and this contributes to the fertilisation event. The dysgenic flies are distinguishable by their mosaic eyes.



This step isolates each reversion event. White eyed F2 male revertants were selected to generate isogenic stable stocks.

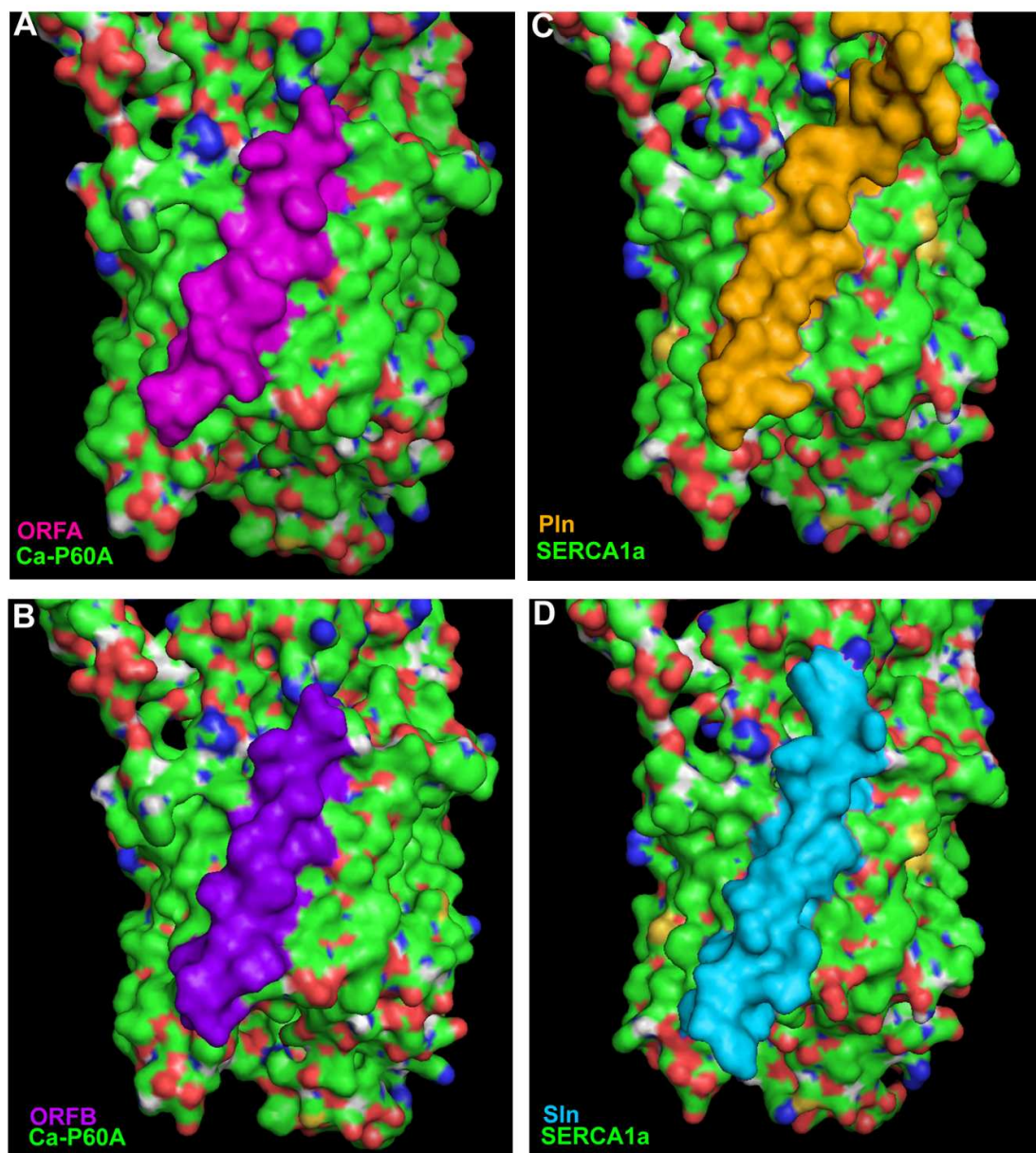




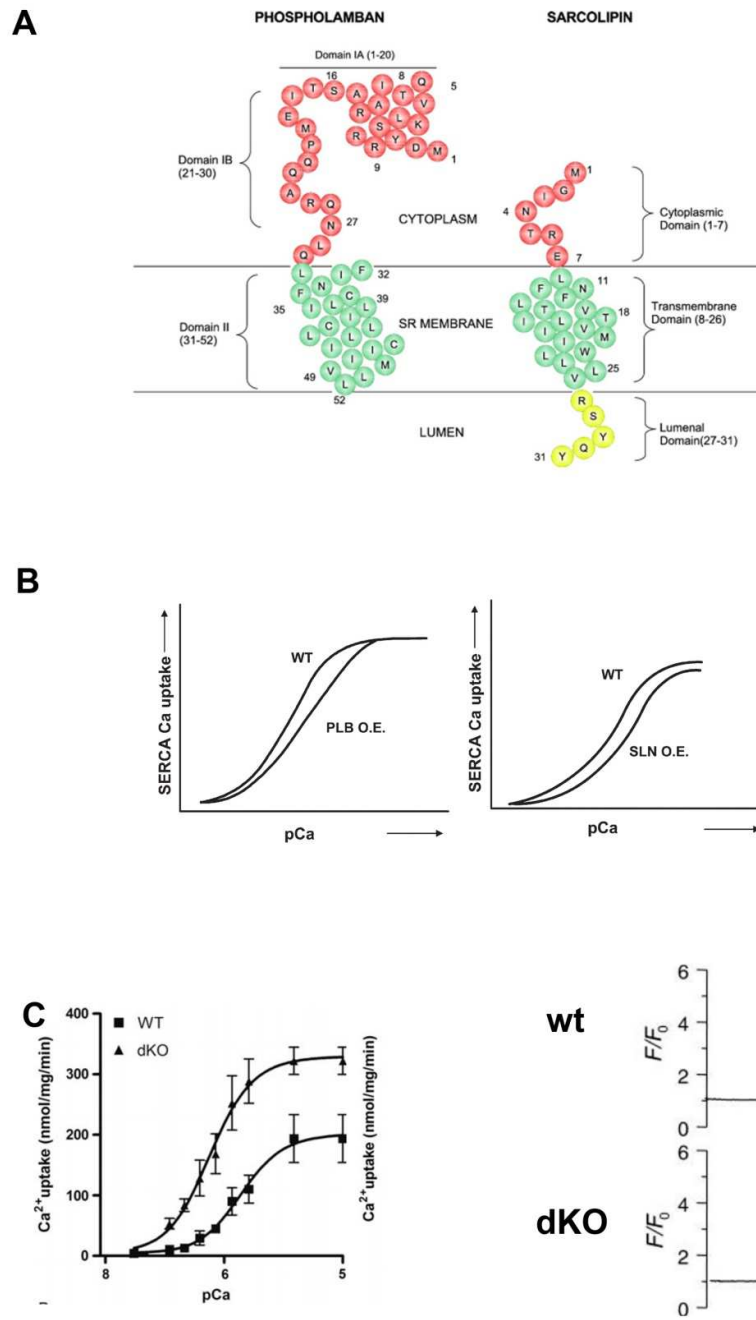
Annex3:DNA alignments of *pncr003;2L*, from the pBac{WH}F02056 lines, and *pncr003;2L* FS construct.

Annex 3(1/2): DNA alignment of the *pncr003;2L RE28911* sequence from *Or-R*, and the pBac{WH} F02056, and pBac{WH} F02056 revertant (RV2) showing that the ORF sequences (pink) are conserved in all these conditions, notice that few other point mutations are present elsewhere in the transcript. The T/C nt substitution in position 361 gives rise to a synonymous L/L mutation in ORF B. The region in blue represents the approximate insertion site of the pBac{WH}F02056 transposon.

Annex 3(2/2): DNA alignment of the *pncr003;2L RE28911* sequences used for the *pncr003;2L* construct and the *pncr003;2L FS* construct. Showing the nt substitutions (red) made to generate frame-shifts in both ORFs (pink).



Annex 4: The Scl peptides can be bioinformatically docked into Ca-P60A, similarly as the vertebrate Sln and Pln peptides into SERCA. (A-D) Molecular models of the Interaction between the *Drosophila* Scl peptides and Ca-P60A, and between the vertebrate Pln and Sln peptides and SERCA1a, courtesy of F.M.G Pearl. For this model the structure of Ca-P60A was modelled on the structure of SERCA1a in its “EI” intermediate structural conformation, using the published crystal structure of SERCA bound to Sarcolipin [102]. (A) Scl ORFA (magenta) and (B)Scl ORFB (purple) dock onto Ca-P60A (green) similarly as (C) Phospholamban (yellow) and (D) Sarcolipin (cyan) dock onto human SERCA1a. Peptide C-termini are facing downwards.



Annex 5: Sln and Pln inhibit the activity of SERCA, and their mutants produce cardiac calcium transients comparable to those of Scl null mutants.

(A) Schematic representation modified from [94], of the homology between the PLB and SLN protein sequence, according to their different domains. Horizontal lines denote the membrane boundaries, and amino acids are shown in circles using their one letter code.

(B) Illustration, modified from [94] , showing the different functional effects of PLB and SLN on SR Ca^{2+} uptake in vertebrate cardiomyocytes: In over expression conditions the inhibitory effect of PLB (PLB O.E) on SERCA calcium uptake is relieved at high calcium concentrations [103] whereas the SLN over-expression (SLN O.E) is inhibitory even at high calcium concentrations [99,107], indicating subtle differences in the mechanism of action of the two regulators.

(C) Effects of the loss of function of the Pln and Sln inhibitors in the activity of SERCA in vertebrate cardiomyocytes, as presented by [101]. Sln and Pln double mutants show an important increase in calcium uptake by SERCA compared to wild-type conditions (Left), The Calcium transient recordings of wild-type (Right, Top) and Sln and Pln double mutant cardiomyocytes (Right, Bottom), show a marked increase in calcium amplitude, very similar to that observed in Scl mutants .

Appendix

Appended to this thesis is the manuscript, as published in the *Science* journal, in which much of the work presented in this thesis was included.



Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames

Emile G. Magny *et al.*

Science **341**, 1116 (2013);

DOI: 10.1126/science.1238802

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 16, 2014):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/content/341/6150/1116.full.html>

Supporting Online Material can be found at:

<http://www.sciencemag.org/content/suppl/2013/08/22/science.1238802.DC1.html>

This article **cites 32 articles**, 13 of which can be accessed free:

<http://www.sciencemag.org/content/341/6150/1116.full.html#ref-list-1>

This article has been **cited by 7 articles** hosted by HighWire Press; see:

<http://www.sciencemag.org/content/341/6150/1116.full.html#related-urls>

This article appears in the following **subject collections**:

Cell Biology

http://www.sciencemag.org/cgi/collection/cell_biol

11. Y. Xiang *et al.*, *Nature* **468**, 921–926 (2010).
12. D. Schmucker, H. Taubert, H. Jäckle, *Neuron* **9**, 1025–1039 (1992).
13. W. B. Grueber, L. Y. Jan, Y. N. Jan, *Development* **129**, 2867–2878 (2002).
14. S. Sanyal, *Gene Expr. Patterns* **9**, 371–380 (2009).
15. C. Montell, *Trends Neurosci.* **35**, 356–363 (2012).
16. E. P. Sawin-McCormack, M. B. Sokolowski, A. R. Campos, *J. Neurogenet.* **10**, 119–135 (1995).
17. O. S. Dominick, J. W. Truman, *J. Exp. Biol.* **117**, 45–68 (1985).
18. T. A. Markow, *Behav. Neural Biol.* **31**, 348–353 (1981).

Acknowledgments: We thank G. Jarretou and B. Polo for technical assistance; F. Casares, J. Bessa, M. Bate, W. Johnson, C. Desplan, F. Pichaud, Y. N. Jan, and stock centers in Kyoto, Bloomington, and Vienna for fly stocks; B. Prud'homme and N. Gompel for wild *D. melanogaster* isolates; L. Vollborn for antibody to PTTH; P. Baroni for light-emitting diode spectrum analysis; and F. Rouyer, J. Simon, S. Sprecher, and laboratory members from O'Connor and Léopold groups for comments on the manuscript. This work was supported by the CNRS, INSERM, Agence Nationale de la Recherche, Fondation pour la Recherche Médicale, European Research Council (grant 268813) to N.M.R., F.A.M., and P.L.; NIH grant K99

HD073239 to N.Y.; Danish Council for Independent Research, Natural Sciences grant 11-105446 to K.F.R.; and NIH grant R01 GM093301 to M.B.O. Materials are available from CNRS under a material transfer agreement.

Supplementary Materials

www.sciencemag.org/cgi/content/full/341/6150/1113/DC1
Materials and Methods
Figs. S1 to S10
References (19–29)

29 May 2013; accepted 9 August 2013
10.1126/science.1241210

Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames

Emile G. Magny,¹ Jose Ignacio Pueyo,¹ Frances M.G. Pearl,^{1,2} Miguel Angel Cespedes,¹ Jeremy E. Niven,¹ Sarah A. Bishop,¹ Juan Pablo Couso^{1*}

Small open reading frames (smORFs) are short DNA sequences that are able to encode small peptides of less than 100 amino acids. Study of these elements has been neglected despite thousands existing in our genomes. We and others previously showed that peptides as short as 11 amino acids are translated and provide essential functions during insect development. Here, we describe two peptides of less than 30 amino acids regulating calcium transport, and hence influencing regular muscle contraction, in the *Drosophila* heart. These peptides seem conserved for more than 550 million years in a range of species from flies to humans, in which they have been implicated in cardiac pathologies. Such conservation suggests that the mechanisms for heart regulation are ancient and that smORFs may be a fundamental genome component that should be studied systematically.

Thousands of small open reading frames (smORFs) exist in animal and plant genomes, yet their relevance and functionality has rarely been addressed because of their challenging properties (1). Detection of small peptides requires specific biochemical and bioinformatics techniques that are rarely used in the characterization of whole genomes. Thus, the number of translated smORFs and their biological functions are still unknown. We and others previously characterized a *Drosophila* gene, *tarsal-less (talpri)*, encoding four smORFs as short as 11 amino acids that are translated and provide essential functions during development (2–4). These results demonstrate that extremely short smORFs can be functional and suggest, when extrapolated by bioinformatics and combined with the latest data from deep RNA sequencing, that hundreds of smORF-encoding transcripts exist in the fly genome (5). However, the *tal* gene is a single example and seems present only in arthropods (2, 3, 6), leaving the questions about the conservation and wider relevance of smORFs unanswered. The characterization of several smORFs displaying conservation of amino-acid sequence, translation, and biological function of the encoded

peptides throughout evolution would be a powerful indicator that smORFs represent an important but neglected part of our genomes.

Using a bioinformatics method (5), we scrutinized the pool of polyadenylated, polysome-associated putative noncoding RNAs in which *tal* was initially included (7) and identified two potentially functional smORFs of 28 and 29 amino acids in the transcript encoded by the gene *putative noncoding RNA 003 in 2L (pncr003:2L)* (Fig. 1A) (6). As with *tal*, these smORFs have similar amino acid sequences to one another and follow strong Kozak sequences (fig. S1A). These peptides are highly hydrophobic, with a predicted alpha-helical secondary structure (fig. S1B).

We corroborated the structure and sequence of the *pncr003:2L* transcript by means of reverse transcription polymerase chain reaction (RT-PCR). Next, we studied *pncr003:2L* expression by means of in situ hybridization, which showed strong expression in somatic muscles and in the post-embryonic heart (Fig. 1, B to D, and fig. S2, A to F). We tested the in vivo translation and subcellular localization of these peptides by generating C-terminal green fluorescent protein (GFP)-tagged fusions within the *pncr003:2L* cDNA of each ORF and expressing these *UAS-smORF-GFP* fusions (fig. S8) in muscles with *Dmef2-Gal4*. We observed the GFP signal at the dyads (Fig. 1E and fig. S1, C and D) (8)—the structures in which

the sarco-endoplasmic reticulum (SER) membrane lies closest to both the plasma membrane and the sarcomeres—in order to facilitate the conversion of the voltage signal into calcium release and muscle contraction (fig. S2G). Similar results were obtained with N-terminal Flag-hemagglutinin-tagged smORFs (*UAS-FH-smORF*) (Fig. 1F and figs. S1, E and F, and S8).

To obtain a null mutant for *pncr003:2L*, we generated two small overlapping deficiencies around the {WH}f02056 insertion (Fig. 1A). Together, these two deletions generate a synthetic homozygous deficiency [{"*Df(2L)scl*"}] eliminating the *pncr003:2L* transcript and the *CG13283* and *CG13282* genes and represents our null condition for the *pncr003:2L* locus, as corroborated with RT-PCR and in situ hybridization (fig. S2, A to F).

Df(2L)scl mutants showed no behavioral or morphological muscle phenotype, even at the ultrastructural level (fig. S2, H to Q). We analyzed muscle function using time-lapse recordings of adult fly hearts (9), which provide an excellent read-out of muscle contraction (Fig. 2A). *Df(2L)scl* mutants showed significantly more arrhythmic cardiac contractions than those of wild-type flies (Fig. 2, A and B; tables S5 and S6; and movies S1 and S2). These effects are due to a requirement for *pncr003:2L* peptides and not the other genes removed in *Df(2L)scl* because the phenotype (i) is mimicked by RNA interference on *pncr003:2L* and (ii) is rescued by restoring expression of *UAS-pncr003:2L* or either of its encoded peptides in *Df(2L)scl* mutants, but is not rescued by smORFs carrying frameshifts in the peptide sequence (Fig. 2B, figs. S3A and S8, and tables S5 and S6). Correspondingly, intracellular electrophysiology recordings in cardiac cells show irregular action potentials (APs), involving “double” and occasionally failed APs in the nonrescued mutants (Fig. 2C, fig. S3C, and table S7).

Because the smORF peptides localize in the dyads, we checked a possible physiological function related to Ca^{2+} trafficking during muscle contraction by visualizing intracellular Ca^{2+} (9). During heart contraction, the Ca^{2+} transients of *pncr003:2L* mutants showed significantly higher amplitudes and steeper decay than those of wild-type controls (Fig. 2D; fig. S3, D and E; and table S8). Overexpression of either peptide in a wild-type fly—but not of frameshifted smORFs—produced reciprocal effects on Ca^{2+} transients but similar arrhythmias to *Df(2L)scl*. Altogether, these results suggest (i) a primary role for the *pncr003:2L* gene during Ca^{2+} trafficking at the SER, which

¹School of Life Sciences, University of Sussex, Falmer, Brighton, East Sussex BN1 9QG, UK. ²Division of Cancer Therapeutics, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK.

*Corresponding author. E-mail: j.p.couso@sussex.ac.uk

Fig. 1. *pncr003:2L* peptide expression in muscles and heart. (A) Annotated genomic region from the Flybase Genome Browser displaying *pncr003:02L*, nearby genes and deficiencies generated in this work, *Df(2L)scl¹²* (green bar), and *Df(2L)scl⁹⁶* (dark blue bar). As transheterozygous, these two deficiencies generate a homozygous deletion (*Df(2L)scl*, red bar), eliminating the *pncr003:2L* transcript and the *CG13283* and *CG13282* genes. (B to D) Expression of *pncr003:2L* mRNA in *Drosophila* muscles (arrowhead), in (B) stage 17 embryos; (C) larval somatic muscles (arrowhead) and heart (arrow), and (D) in the adult heart (arrow). (E to E'') ORFA-GFP expression (green; arrowheads) surrounding the phalloidin-stained sarcomeres (magenta) in adult transversal heart fibers. (F to F'') FH-ORFA peptides display a reticular pattern (green; arrowheads) in adult longitudinal heart fibers labeled with phalloidin (magenta). Blue, 4',6-diamidino-2-phenylindole (DAPI)-stained nuclei.

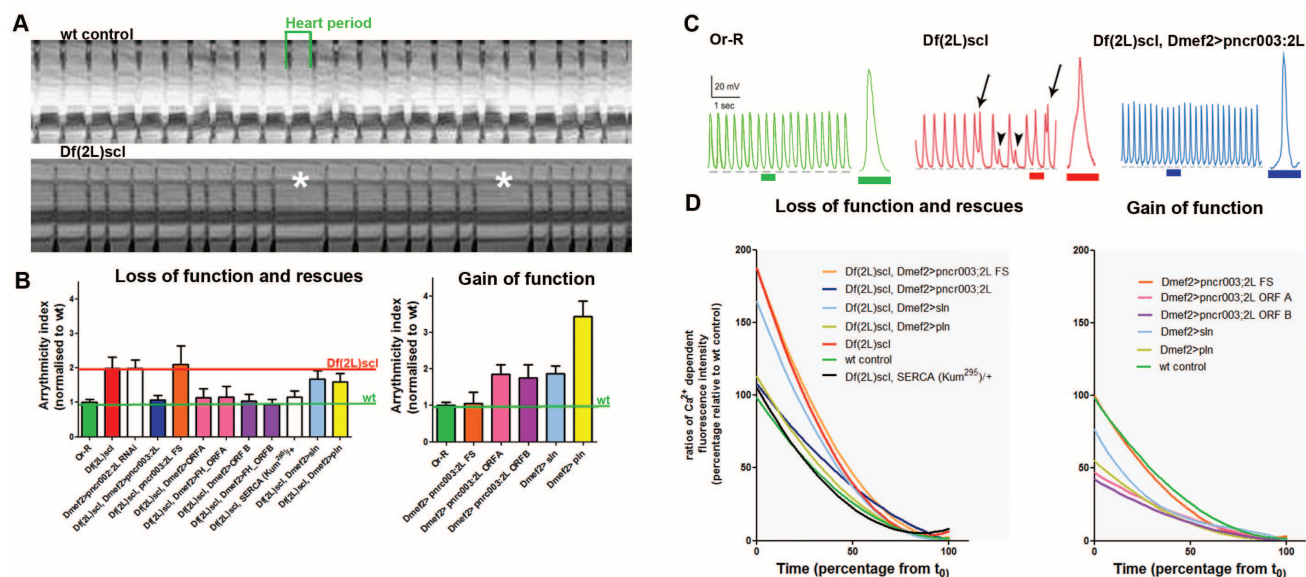
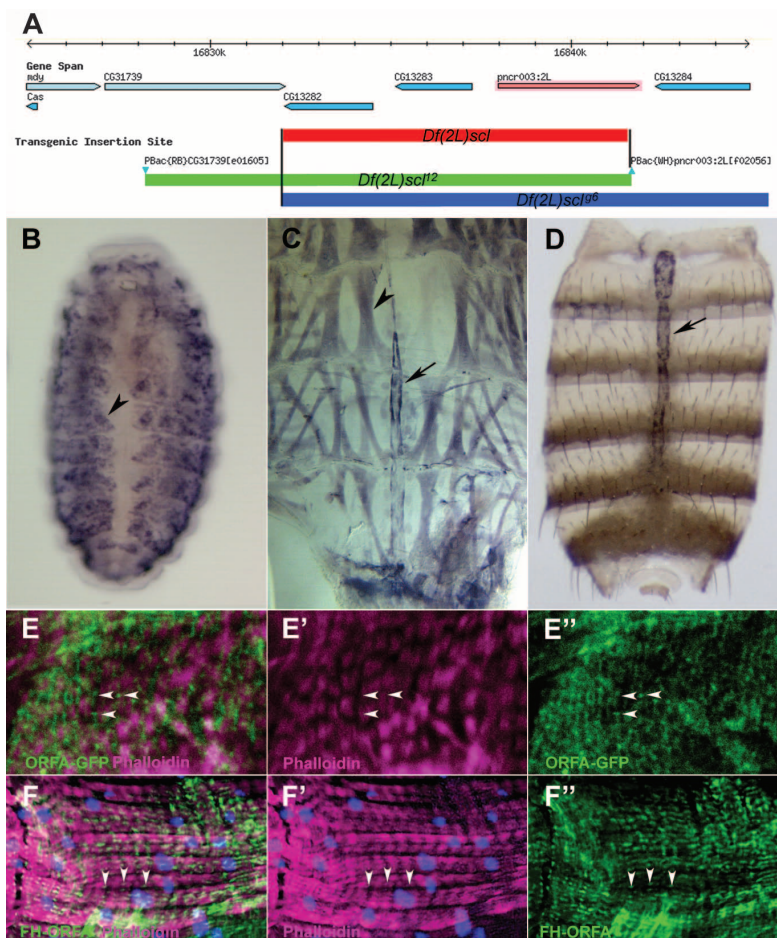
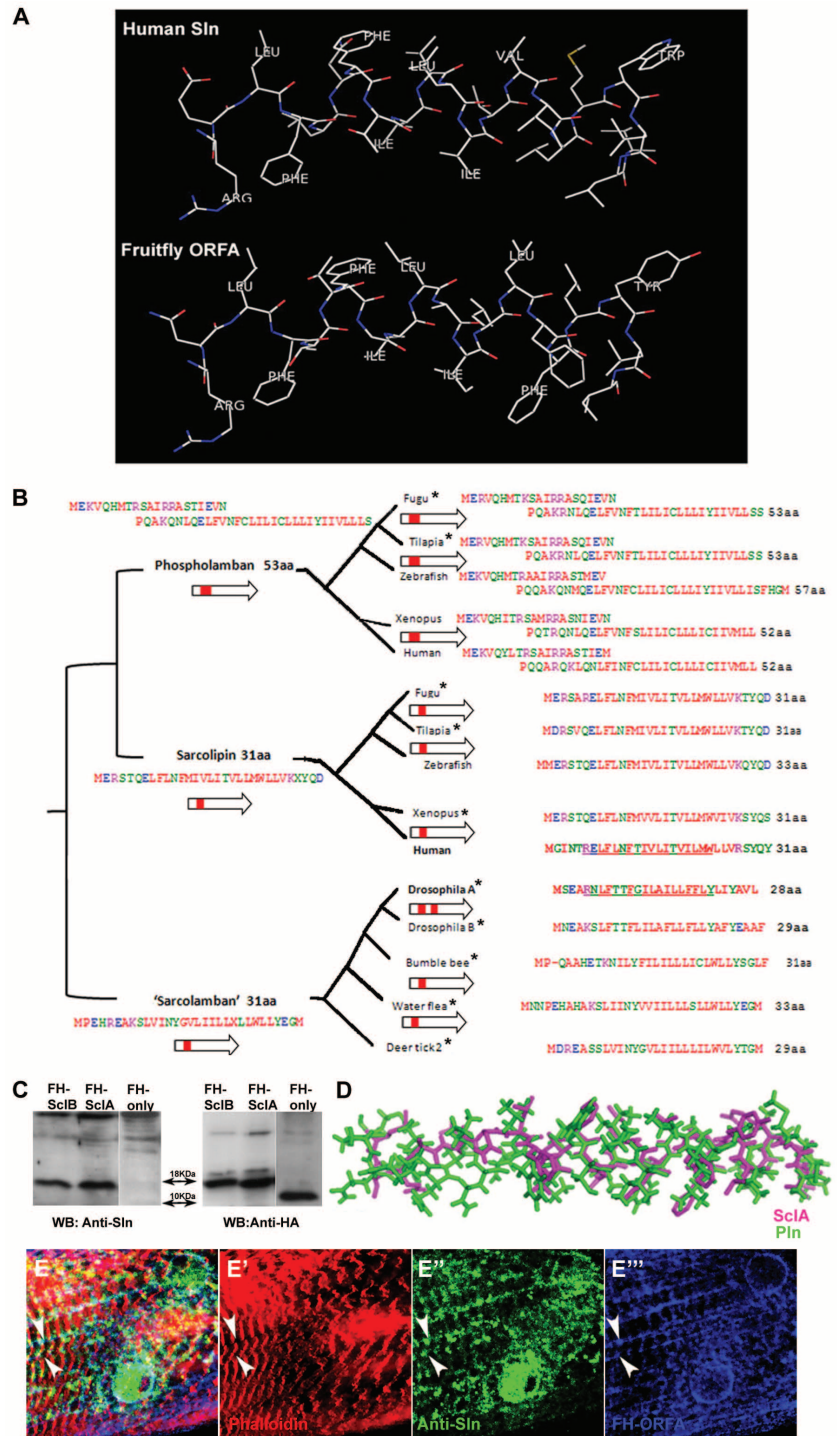


Fig. 2. Role of *pncr003:2L* in cardiac muscle contraction. (A) Kymographs comparing the pattern of heart contractions for wild-type and *Df(2L)scl* hearts. The mutant shows irregular periods, some being abnormally long (asterisk). A normal heart period is indicated (green). (B) Arrhythmicity index of *pncr003:2L* loss-of-function and rescue genotypes (left) and excess of function genotypes (right), normalized to age matched wild-type controls (9). Columns represent mean, and error bars represent SE. (C) Sample traces of intracellular recordings from adult cardiomyocytes of wild-type (green); *Df(2L)scl* (red); and *Df(2L)scl* rescued by *UAS-pncr003:2L* (blue). Arrows indicate "double" action

potentials. Arrowheads indicate failed action potentials. Gray dashed line indicates resting potentials. Sample peaks from each trace (underlined) appear magnified. (D) Ca^{2+} transients during heart contraction of *Df(2L)scl* and rescue genotypes (left) and gain-of-function genotypes (right) color-coded as in (B). The fluorescent Ca^{2+} sensor G-CaMP3 was used to visualize calcium levels. Y axis values are ratios of calcium dependent fluorescence on its decay phase normalized to basal intensities and presented as percentages relative to wild-type controls; x axis values are percentage of time from the point of maximum transient amplitude.

Fig. 3. Putative homology of sequence and structure between human and *Drosophila* peptides. (A) Secondary structure of the conserved domain [underlined in (B)] of Sarcolipin (top) and *Drosophila pncr003:2L*ORFA peptide (bottom). Blue, nitrogen atoms; red, oxygen atoms. (B) Phylogenetic tree of vertebrate and arthropod (*pncr003:2L*, labeled "Sarcolamban") peptides. Asterisks indicate sequences identified in this study (supplementary data file S1). Putative ancestral consensus sequences (left) and further analysis (fig. S4) (9) suggest that the two vertebrate peptides arose from a duplication of a single ancestor that also diverged independently into the different arthropod Sarcolamban peptides. Analysis of RNA (cDNA) sequences (arrows) indicates that all peptides arise from single smORFs (red boxes) uninterrupted by exons, suggesting that ancestral peptides were also encoded by smORFs. (C) Western blots from *Drosophila* S2 cells showing that the antibody to human Sarcolipin (left lanes) recognizes the *Drosophila* FH-tagged Sarcolamban18-kD peptides SclA and SclB, but not the 10-kD FH-tag alone. Right lanes show positive controls, with antibody to HA recognizing all peptides. (D) A compatible structure for Sarcolamban-A (magenta) is obtained by threading it onto the C-terminal domain of vertebrate Phospholamban (green). (E to E''') *Drosophila* FH-SclA peptides (arrowheads) surrounding the sarcomeres (red) are recognized by antibodies to Sarcolipin (green) and Flag (blue) in larval somatic muscles.



would be secondarily required for regular muscle contraction; and (ii) that such a role is mediated by the peptides encoded by the 28- and 29-amino acid smORFs.

We searched for conservation of these smORFs in other species by using Basic Local Alignment Search Tool (BLAST) and only identified them in other Drosophilids [with K_a/K_s scores of <0.2 supporting translation (10)]. Because the *pncr003:2L* peptides have a predicted helical structure, we searched for possible structural homologs (9) and retrieved the 30-amino acid human *sarcolipin*

(Sln) peptide (Fig. 3A and tables S1 and S2) (11). However, the Sln and *pncr003:2L* peptides display noticeable differences in their amino acid sequences (Fig. 3B). If they were true homologs, peptides with intermediate sequences should exist in the stem lineages to both flies and humans. We devised a bioinformatics protocol (9) to identify possible *pncr003:2L* homologs in arthropods (Fig. 3B and fig. S4) plus nonannotated homologs of *sln* and its longer paralogue *phospholamban* (*pln*) (Fig. 3B and fig. S4) (12), until basal arthropod smORFs identified basal vertebrate homologs with

the expected intermediate amino acid changes (fig. S4, A to C). Supporting their putative homology, we found that (i) antibodies to sarcolipin recognize the *pncr003:2L* peptides (Fig. 3, C and E, and fig. S5, A and B), and (ii) threading the *pncr003:2L* amino acid sequences on the Pln three-dimensional (3D) structure (13) also produces a compatible structure (Fig. 3D and tables S1 and S2).

A phylogenetic tree of all these peptides suggests that Sln and Pln emerged from a gene duplication in vertebrates, whereas an independent and more recent duplication in flies gave rise to

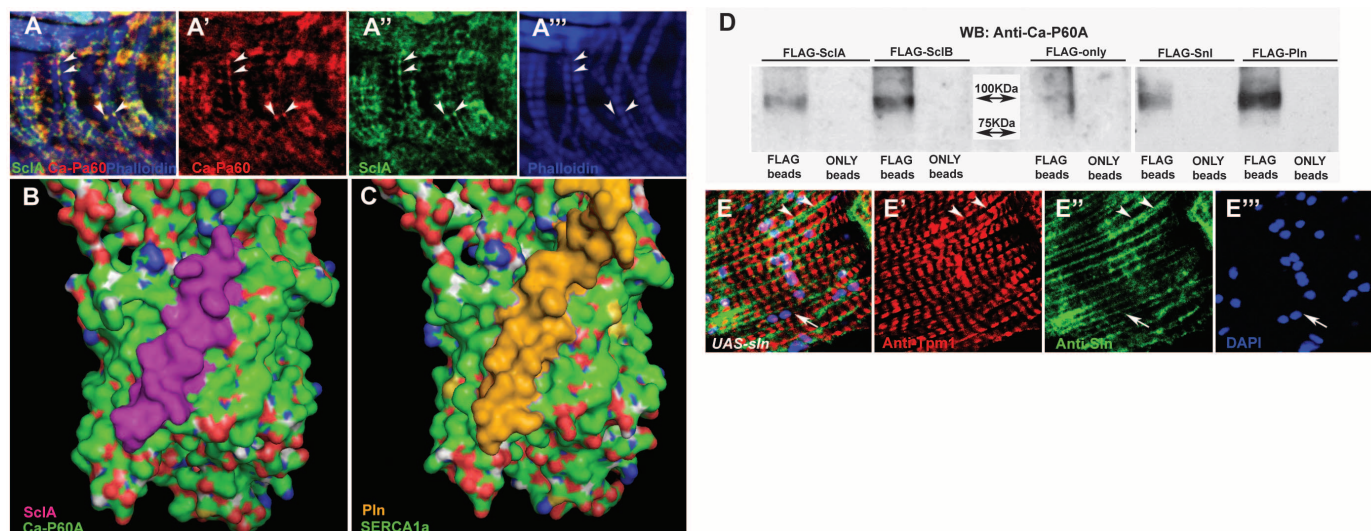


Fig. 4. Sarcolamban interacts with Ca-P60A SERCA. (A to A'') Co-localization of sarcolamban FH-SclA peptides (green) and Ca-P60A SERCA (red) in the SER and dyads (arrowheads) surrounding the adult heart sarcomeres (blue, phalloidin). (B and C) Interaction between the *Drosophila* SclA (magenta) and Ca-P60A, modeled from vertebrate SERCA1a in the EI conformation (9). SclA docks onto Ca-P60A similarly as Phospholamban (yellow) and Sarcolipin (fig. S5, D and E) onto human SERCA1a (C). Peptide C-termini

are down. (D) FH-tagged *Drosophila* SclA and SclB and the human Sln and Pln peptides pull-down the 100-kD *Drosophila* Ca-P60A (revealed with antibody to Ca-P60) from transfected S2 cells. Negative control lanes with Flag-only peptides or beads without antibodies ("ONLY beads") do not show similar Ca-P60A signal. (E to E'') Human Sln peptides (green; arrowheads) expressed in the *Drosophila* adult heart surround the sarcomeres (red; labeled with antibody to Tropomyosin1). Blue, DAPI-stained nuclei (arrow).

pncr003:2L ORFA and ORFB peptides. The tree, sequence alignments, and further bioinformatics analysis (fig. S4, supplementary data file S1, and tables S1 and S2) (9) are altogether compatible with a single origin for the Sln, Pln, and *pncr003:2L* peptides from an ancestral peptide-encoding smORF of ~30 amino acids (Fig. 3B and fig. S4B). We suggest that *pncr003:2L* and its arthropod homologs should be renamed *sarcolamban* (*scl*) in order to reflect their similarity and probable homology to vertebrate *sln* and *pln*.

Conservation of smORFs across such an evolutionary distance (>550 million years of divergence) has not been described; therefore, we scrutinized their functional homology. Sln and Pln regulate Ca^{2+} traffic in mammal muscles by dampening the activity of the Sarco-endoplasmic Reticulum Ca^{2+} adenosine triphosphatase (SERCA), whose function is to retrieve Ca^{2+} from the cytoplasm back into the SER, leading to muscle relaxation (fig. S2G) (14). The effects of removing *sln* upon the vertebrate muscle Ca^{2+} transients are remarkably similar to the effects we observed in *Df(2L)scl* mutants (Fig. 2D) (15). Furthermore, abnormal levels of Sln expression have been related to human heart arrhythmias (16), and Sln and Pln have been shown to bind SERCA (17). In flies, the Scl peptides colocalize with *Drosophila* SERCA (Ca-P60A) (Fig. 4A and fig. S5C) and coimmunoprecipitate with it (Fig. 4D). Furthermore, the arrhythmia and abnormal transients of *Df(2L)scl* mutants are corrected by reducing the function of *Ca-P60A* (Fig. 2, B to D), a genetic interaction that is consistent with a down-regulating role of Scl upon SERCA activity (18). Last, threading the sequence of Ca-P60A onto the

3D structure of vertebrate SERCA produces a compatible structure that seems able to dock Scl similarly to Sln and Pln binding to SERCA (Fig. 4, B and C; fig. S5, D and E; and tables S3 and S4) (17).

Our studies suggest that Sln and Pln can bind fly Ca-P60A and can resemble Scl function. Modeling suggests that fly and vertebrate peptides could bind each other's SERCA (tables S3 and S4), and indeed human peptides can pull down fly Ca-P60A (Fig. 4D). Sln and Pln expressed in fly muscles and cultured cells localize similarly to Scl and Ca-P60A (Fig. 4E and figs. S5F and S6) and produce arrhythmias and Ca^{2+} transients similar to those produced by overexpressing fly Scl peptides (Fig. 2, B and D). Furthermore, expression of human Pln in *Df(2L)scl* flies can rescue the mutant Ca^{2+} transients toward wild type, and the strong arrhythmia phenotype of ectopic Pln is itself reduced (Fig. 2, B and D). The human peptide overexpression and rescue effects do not completely reproduce those observed with fly peptides, and this suggests that although this family of peptides may share a regulatory function on Ca^{2+} pumps, each seems finely tuned to its own species-specific SERCA regulation.

Altogether, our results suggest that this family of peptides may represent an ancient system for the regulation of Ca^{2+} traffic, whose alteration can result in irregular muscle contractions. We propose that the *Drosophila sarcolamban* (*scl*) gene, previously annotated as the long noncoding RNA *pncr003:2L*, actually encodes two functional smORFs of 28 and 29 amino acids that are translated into bioactive peptides. The analysis of related amino acid sequences across multiple species is compatible with a conservation of these

peptides and their putative molecular structure from flies to vertebrates, correlated with the conservation of their biological role in regulating Ca^{2+} uptake at the SER. We speculate that this remarkable conservation, together with previous reports on the *tal* gene (2–4), might indicate that smORFs can reveal both sequence conservation and important biological functions. Bioinformatics predictions (1, 5) and recent ribosomal profiling data from vertebrates (19) suggest that translated smORFs may be abundant. We believe that smORFs cannot be dismissed as irrelevant, but that their functionality should be considered whenever encountered.

References and Notes

1. M. A. Basrai, P. Hieter, J. D. Boeke, *Genome Res.* **7**, 768–771 (1997).
2. M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso, *PLoS Biol.* **5**, e106 (2007).
3. J. I. Pueyo, J. P. Couso, *Dev. Biol.* **355**, 183–193 (2011).
4. T. Kondo et al., *Science* **329**, 336–339 (2010).
5. E. Ladoukakis, V. Pereira, E. G. Magny, A. Eyre-Walker, J. P. Couso, *Genome Biol.* **12**, R118 (2011).
6. J. Savard, H. Marques-Souza, M. Aranda, D. Tautz, *Cell* **126**, 559–569 (2006).
7. J. L. Tupy et al., *Proc. Natl. Acad. Sci. U.S.A.* **102**, 5495–5500 (2005).
8. R. I. Razzaq A et al., *Genes Dev.* **15**, 22 (2001).
9. Material and methods are available as supplementary materials on Science Online.
10. A. Nekrutenko, K. D. Makova, W. H. Li, *Genome Res.* **12**, 198–202 (2002).
11. A. Wawrzynow et al., *Arch. Biochem. Biophys.* **298**, 620–623 (1992).
12. P. Bhupathy, G. J. Babu, M. Periasamy, *J. Mol. Cell. Cardiol.* **42**, 903–911 (2007).
13. K. Okenoid, J. J. Chou, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 10870–10875 (2005).

14. M. Periasamy, A. Kalyanasundaram, *Muscle Nerve* **35**, 430–442 (2007).
15. G. J. Babu et al., *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17867–17872 (2007).
16. M. Shanmugam et al., *Biochem. Biophys. Res. Commun.* **410**, 97–101 (2011).
17. M. Asahi et al., *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5040–5045 (2003).
18. G. J. Babu et al., *J. Biol. Chem.* **281**, 3972–3979 (2006).
19. N. T. Ingolia, L. F. Lareau, J. S. Weissman, *Cell* **147**, 789–802 (2011).

Acknowledgments: We thank Rose Phillips, Roger Phillips, and J. Thorpe for technical support; M. Ramaswami for the antibody to Ca-P60A; and F. Casares, M. Baylies, I. Galindo, C. Alonso, and laboratory members for manuscript comments. E.M. was supported by Conacyt, F.P. was supported by a Daphne Jackson Fellowship and the UK Medical Research Council, and J.N. was supported by a Royal Society University Fellowship. Otherwise, this work was funded by a Wellcome Trust Fellowship (ref 087516) awarded to J.P.C. The GenBank accession number for *Drosophila* Scl sequences is NR_001662.

Supplementary Materials

www.sciencemag.org/cgi/content/full/341/6150/1116/DC1
Materials and Methods
Figs. S1 to S8
Tables S1 to S9
References (20–34)
Movies S1 and S2
Supplementary data file S1

5 April 2013; accepted 7 August 2013
10.1126/science.1238802

A Causative Link Between Inner Ear Defects and Long-Term Striatal Dysfunction

Michelle W. Antoine,¹ Christian A. Hübner,² Joseph C. Arezzo,¹ Jean M. Hébert^{1,3*}

There is a high prevalence of behavioral disorders that feature hyperactivity in individuals with severe inner ear dysfunction. What remains unknown is whether inner ear dysfunction can alter the brain to promote pathological behavior. Using molecular and behavioral assessments of mice that carry null or tissue-specific mutations of *Slc12a2*, we found that inner ear dysfunction causes motor hyperactivity by increasing in the nucleus accumbens the levels of phosphorylated adenosine 3',5'-monophosphate response element-binding protein (pCREB) and phosphorylated extracellular signal-regulated kinase (pERK), key mediators of neurotransmitter signaling and plasticity. Hyperactivity was remedied by local administration of the pERK inhibitor SL327. These findings reveal that a sensory impairment, such as inner ear dysfunction, can induce specific molecular changes in the brain that cause maladaptive behaviors, such as hyperactivity, that have been traditionally considered exclusively of cerebral origin.

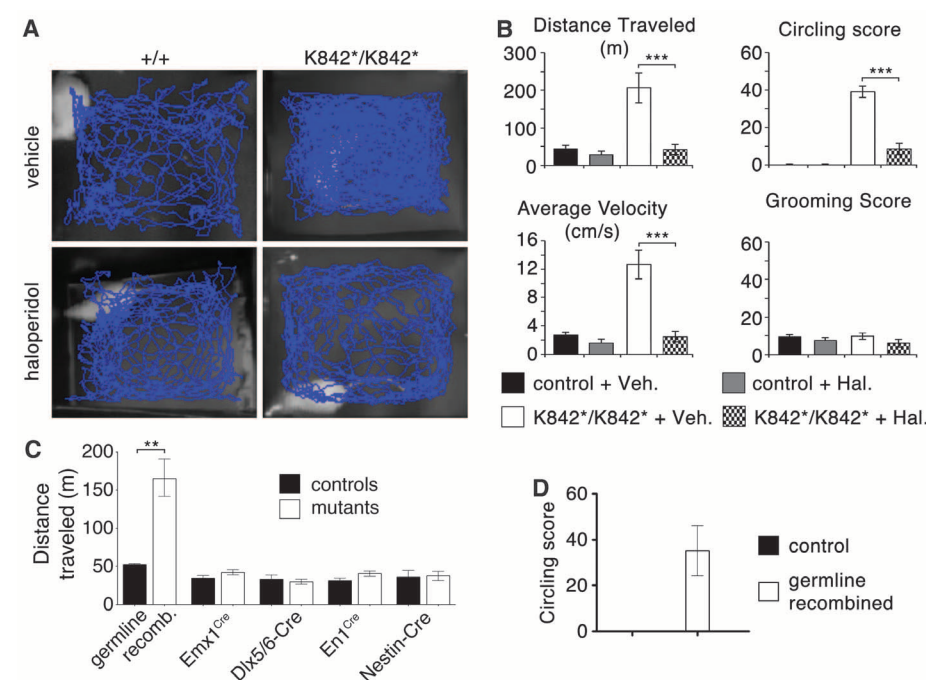
The inner ear contains the cochlea, devoted to hearing, and the vestibular end organs, dedicated to balance. In 20 to 95% of children with severe hearing loss, auditory and vestibular dysfunction occur concurrently (1, 2). In such cases, there is a high incidence of behavioral

disorders that feature hyperactivity as a core diagnostic symptom (3–5). Although socioenvironmental variables have been proposed as risk factors (6), it is unclear whether sensory impairments, such as inner ear defects, can directly induce specific changes in the brain that lead to

maladaptive behavior. In nonhuman vertebrates, including rodents and frogs, surgical or pharmacological lesions to the vestibulo-auditory system are also linked to long-term changes in locomotor activity, although, to date, the associations between ear dysfunction and behavior remain unexplained (7–9). Genetic mouse models of inner ear dysfunction can exhibit increased levels of locomotor hyperactivity (10), but because the gene is mutated in the brain, as well as the inner ear, the causal neural underpinnings of this behavior remain unknown.

Slc12a2 (also known as *Nkcc1*) is a gene that encodes a sodium-potassium-chloride cotransporter broadly expressed in tissues, including the inner ear and central nervous system (CNS) (11, 12). The *Slc12a2* mutant mice used in this study arose spontaneously in our mouse colony and exhibit increased levels of motor hyperactivity, including locomotion, circling, and head tossing (Fig. 1, A and B; movie S1; and fig. S1A),

Fig. 1. *Slc12a2*^{K842*/K842*} mutants display a dopamine receptor-mediated increase in locomotor activity that cannot be explained by disruption of *Slc12a2* in the brain. (A) Traces and (B) quantification of mouse locomotion in an open field showing that haloperidol alleviates locomotor activity and circling in *Slc12a2*^{K842*/K842*} mice without affecting grooming [****P* < 0.0001; repeated measures analysis of variance (ANOVA) with Bonferroni post hoc comparison]. (C) Germline recombination of *Slc12a2*^{flx/flx} mice recapitulates the increased locomotion of the *Slc12a2*^{K842*/K842*} mutant (*P* = 0.0032, unpaired two-tailed test). Mice lacking *Slc12a2* in the neocortex and hippocampus (*Emx1*^{Cre/+}; *Slc12a2*^{flx/rec}), striatum (*Dlx5/6-Cre*; *Slc12a2*^{flx/flx}), cerebellum (*En1*^{Cre/+}; *Slc12a2*^{flx/flx}), and CNS (*Nestin-Cre*; *Slc12a2*^{flx/flx}) display normal levels of motor activity (unpaired two-tailed test). (D) Germline recombination of the *Slc12a2*^{K842*/K842*} mutants. *n* = 4 to 11 mice per genotype. All data are means ± SEM.



¹Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY 10461, USA. ²Jena University Hospital, Institute of Human Genetics, Jena 07743, Germany. ³Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA.

*Corresponding author. E-mail: jean.hebert@einstein.yu.edu